இயந்திரமொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும்

Machine translation and preparation of tools for machine translation

இராசேந்திரன் சங்கரவேலாயுதன்

rajushush@gmail.com

தஞ்சாவூர்

2010

என்னுரை

பல ஆண்டுகளாக எனது கணிப்பொறியில் உறங்கிக் கொண்டிருந்த "இயந்திரமொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும் (Machine translation and preparation of tools for machine translation)" என்ற இவ்வேட்டிற்கு தற்போது உயிர் அளித்துள்ளேன். இதைச் செம்மைப்படுத்தி நூலாக வெளியிடவேண்டும் என்பது எனது அவா. காலம் கடந்து கொண்டிருக்கின்றது. பணி ஓய்வுக்குப் பின்னர் இதைச் செம்மைப்படுத்தி நூலாக வெளியிடவேண்டும் என்று எண்ணியிருந்தேன். தஞ்சைத் தமிழ்ப்பல்கலைக் கழகத்தில் 2011 ஜூனில் ஓய்வு பெற்ற உடன் அவ்வாண்டு ஜூலையிலேயே கோயம்புத்தூர் அமிர்தா பல்கலைக் கழகத்தில் பணிக்குச் சேர்ந்து பணியாற்ற வேண்டிய சூழல். எனவே இவ்வேட்டைத் திருத்தக் காலம் கிடைக்கவில்லை. எனவே திருத்தமின்றி அவ்வாறே தருகின்றேன்.

தமிழ் நாட்டில் மொழியியல் துறைகள் நலிந்து கொண்டிருக்கின்றன. மொழியியல் படிப்பதற்கு மாணவர்கள் தயாராய் இல்லை. ஏனென்றால் மொழியியல் கல்லூரிகளில் பாடமாக அமையாததால் வேலை வாய்ப்பு குறைவு அல்லது இல்லை. மொழியியலின் இன்றியமையாமை மாணவர்களுக்கும் மொழி அறிஞர்களுக்கும் போய்ச்சேரவில்லை. இளங்கலை நிலையிலும் முதுகலை நிலையிலும் மொழிப் பாடங்களுடன் மொழியியல் கற்பிக்கப்படுவதில்லை. இத்தகைய சூழலில் இவ்வேடு தேவையற்றதாக இருக்கலாம். இருப்பினும் கணினி உதவியால் மொழி

ஆய்வு செய்ய விரும்பும் தமிழ் மட்டும் அறிந்தவர்களுக்கு இது பயன்படலாம் என்ற நோக்கில் இதை இங்கு சமர்ப்பிக்கின்றேன். ஆங்கிலம் அறியாமல் தமிழ் மட்டும் அறியும் மாணவர்களுக்கு இது பயன்படக்கூடும்.

அன்புடன்

இராசேந்திரன் சங்கரவேலாயுதன்

பொருளடக்கம்

வரிசை எண்	தலைப்புகள்	பக்கம்
1	இயல் – 1: முன்னுரை	14
2	இயல் – 2: இயந்திர மொழிபெயர்ப்பு	17
2.1	அறிமுகம்	17
2.2	கணிப்பொறி வழிமொழிபெயர்ப்பின் தேவை	25
2.3	கணிப்பொறி வழி மொழிபெயர்ப்பின் வகைகள்	26
2.3.1	கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு	26
2.3.2	மனித உதவியுடன் கணிப்பொறி மொழிபெயர்ப்பு	27
2.3.3	முற்றிலும் தானியங்கி மொழிபெயர்ப்பு	27
2.4	கணிப்பொறி வழி மொழிபெயர்ப்பின் சில முயற்சிகள்	28
2.5	கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்	30
2.6	கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்	32
2.7	இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி	38
2.7.1	இடைமொழி அணுகுமுறை	39
2.7.2	சொல் நிலையில் மாற்றத்தைப் பயன்படுத்திய நேரடி அணுகுமுறை [அனுசாராத் திட்டம்]	41
2.7.3	ஆங்கிலத்திலிந்து இந்திய மொழிகளுக்கு இயந்திர உதவியுடன் மொழிபெயர்ப்பு [ஆங்கில பாரதி]	43
2.8	தமிழ்சார் இயந்திர மொழிபெயர்ப்பு	52
2.8.1	தமிழ்சார் இயந்திர மொழிபெயர்ப்பு முயற்சிகள்	**51**
2.8.1.1	தமிழ் உருஷ்யன் மொழிபெயர்ப்புத்திட்டம்	**51**
2.8.1.2	உலக வலைப்பின்னல் மொழி – தமிழுக்கான இடைமொழி	**51**

	இயந்திர மொழிபெயர்ப்பு	
2.8.1.3	ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்கும் திட்டம்	53
2.8.2	தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்	55
2.8.2.1	உருபனியல் பகுப்பாய்வு	55
2.8.2.2	சொல்வகைப்பாட்டு அடையாளப்படுத்தல்	56
2.8.2.3	தொடரியல் பகுப்பாய்வு	57
2.8.2.4	பொருண்மையியல் பகுப்பாய்வு	58
2.8.2.5	தொடரியல் சார் மாற்றம்	58
2.8.2.6	அகராதி பொருத்தம்	59
2.8.2.7	உருபனியல் உருவாக்கம்	59
2.8.3	தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்	60
2.8.3.1	கருவிகள்	60
2.8.3.1.1	உருபனியல் பகுப்பாய்வி	60
2.8.3.1.2	அடையாளப்படுத்தி	60
2.8.3.2.3	பகுப்பான்	61
2.8.3.2.5	பொருண்மையியல் பகுப்பாய்வி	61
2.8.3.2.6	உருபனியல் உருவாக்கி	61
2.8.3.2	மூலவளங்கள் அகராதி	62
2.8.3.2.1	அகராதி	62
2.8.3.2.2	மொழி கடந்த அகராதி	62
2.8.3.2.3	தரவுத் தொகுதி	62
2.8.3.2.4	மாற்றமைவு இலக்கணம்	64
2.8.3.2.5	இணைவமைதி அகராதி	64
2.9	நிறைவுரை	64

3	இயல் – 3: தமிழுக்குத் தரவுத் தொகுதிகளை உருவாக்குதல்	65
3.1	அறிமுகவுரை	65
3.2	ஆங்கில மொழிக்கான தரவுத் தொகுதிகள்/ தரவுத்தொகுதிகள்	67
3.2.1	பிரவுன் தரவுத் தொகுதி	67
3.2.2	லோப் தரவுத் தொகுதி	68
3.2.3	ஆங்கிலத்தில் ஆஸ்டிரேலியன் தரவுத்தொகுதி	68
3.2.4	எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி	69
3.2.5	லோப் தரவுத்தொகுதி	70
3.2.6	பிரிட்டிஷ் தேசிய தரவுத்தொகுதி	71
3.2.7	அமெரிக்க தேசிய தரவுத்தொகுதி	72
3.2.8	ஆங்கில வங்கி	72
3.3	இந்திய மொழிகளுக்கான தரவுத்தொகுதிகள்	73
3.3.1	இந்திய மொழிகளின் MIT தரவுத்தொகுதி	73
3.3.2	இந்திய மொழிகளுக்கான தரவுத்தொகுதிகள்	74
3.3.3	EMILLE – தரவுத்தொகுதி	76
3.3.4	EMILLE மற்றும் CIIL ஒரு எழுத்து தரவுத்தொகுதி	77
3.3.5	தமிழுக்கான தரவுத்தொகுதி தயாரித்தல்	77
3.4	உரை தரவுத் தொகுதி ஆய்வு	79
3.4.1	நிகழ்வெண் ஆய்வு	80
3.4.2	சொல்வருகை ஆய்வி	80
3.4.3	சூழலில் முக்கியச்சொல்	81
3.4.4	சொல் பகுப்பாய்வு	83
3.4.5	திரிபுறாத சொற்களின் ஆய்வு	83
3.4.6	திரிபுற்ற சொற்களைப் பகுத்தாய்தல்	85

3.4.7	இரட்டை சொற்களைப் பகுத்தாய்தல்	86
3.5	தரவுத் தொகுதி அடையாளப் படுத்தல்	86
3.5.1	சொல் அடையாளப்படுத்தல்	86
3.5.2	இலக்கணம் அடையாளப்படுத்தல்	87
3.5.3	சொல்பொருள் அடையாளப்படுத்தல்	87
3.6	மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதி	87
3.6.1	மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதியின் முக்கியத்துவம்	88
3.6.2	மொழித்தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்	90
3.7	மொழிபெயர்ப்புக்கு உதவும் ஒழுங்கு முறைகளுக்கு மூலவளமாகத் தரவுத்தொகுதி	91
3.8	மனித- இயந்திர இடைமுக ஒழுங்கு முறைகளுக்கு மூலவளமாகத் தரவுத்தொகுதி	91
3.9	பேச்சுத் தொழில் நுட்பத்தில் தரவுத்தொகுதி	92
3.10	இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி	92
3.10.1	நோக்கம்	95
3.10.2	வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்	96
3.10.3	தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை	97
3.10.4	தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்	99
3.10.5	மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கம்	99
3.10.6	மொழிபெயர்ப்பு தரவுத்தொகுதியைப் பொருத்தமாக வரிசைப்படுத்தல்	101
3.10.7	இந்திய மொழிகளின் இன்றைய நிலை	103
3.10.8	உருவாக்கம்	103
3.10.9	மொழிபெயர்ப்பு தரவுத்தொகுதிகளில் மொழியின்	104

	செயல்பாடுகள்	
3.10.10	மொழிபெயர்ப்பு ஆய்வு	105
3.10.11	இருமொழிய அகராதியின் உருவாக்கம்	106
3.10.12	மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு	106
3.10.13	கலைச்சொல் தகவல் வங்கியின் உருவாக்கம்	107
3.10.14	சொல்தேர்வுக் கட்டுப்பாடு	108
3.10.15	சொல் மயக்கத்தை நீக்குதல்	109
3.10.16	இலக்கணப் பொருத்தம்	111
4	இயல் – 4: தமிழ்த் தரவுத் தொகுதியைச் சொல்வகைப் பாட்டிற்கு அடையாளப்படுத்தல்	112
4.1	அறிமுகம்	112
4.2	தரவுத்தொகுதியை அடையாளப்படுத்தல்	112
4.2.1	சொல்வகைப்பாடு அடையாளப்படுத்தல்	113
4.2.2	முன்வரு கிளவி அடையாளப்படுத்தல்	114
4.2.3	மீக்கூறு அடையாளப்படுத்தல்	115
4.2.4	பொருண்மை அடையாளப்படுத்தல்	116
4.2.5	கருத்தாடல் அடையாளப்படுத்தல்	117
4.2.6	பகுத்துக்குறித்தல்	117
4.2.7	தலைச்சொல்லாக்கம்	119
4.3	சொல்வகை அடையாளப்படுத்தலின் முக்கியத்துவம்	120
4.4	சொல்வைப்பாட்டை அடையாளப்படுத்தலின் வகைகள்	122
4.4.1	விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்	122
4.4.2	புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்	123

4.4.3	மாற்றம் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்	125
4.5	சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம்	128
4.5.1	சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் பற்றிய முந்தைய முயற்சிகள்	128
4.5.1.1	தொடக்கக்காலச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம் [1980-க்கு முன்]	128
4.5.1.2	ஆங்கில சொல் வகைப்பாட்டு குழுமங்கள் [1980-க்கு பிறகு]	128
4.5.2	அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு	129
4.5.3	ENGLES வழிக்காட்டல்களின் அடிப்படையில் உருவாக்கப்பட்ட சில சமீபக் காலத்திய சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்	132
4.6	தமிழ்த் தரவுத் தொகுதிகளை சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்துதல்	132
4.6.1	தமிழ்ச் சொல்வகைப்பாடுகள்	132
4.6.2	தமிழுக்கான சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள்	151
4.6.2.1	CIIL நிறுவனத்தின் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்	151
4.6.2.2	AU-KBC நிறுவனத்தின் தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்	153
4.6.2.3	அமிர்தா தொழில்நுட்பக் கல்லூரி பயன்படுத்திய தமிழுக்கானச் சொல்வகை அடையாளக் குழுமம்	158
4.6.2.4	IIITஹைதராபாத்தில் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமம்	161

4.6.2.5	AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு	162
4.6.2.6	பிற சொல்வகை அடையாளப்படுத்திகள்	165
4.6.2.6.1.	வாசு அரங்கநாதனின் டேக் தமிழ்	166
4.6.2.6.2.	கணேசனின் சொல்வகை அடையாளப்படுத்தி	166
4.6.2.6.3.	RCILTS-இன் தமிழ் கதம்பம்	167
4.7.	இந்தி மொழியிலிருந்து இந்திய மொழி இயந்திர மொழிப்பெயர்ப்பு ஒழுங்கு முறையில் சொல் வகைப்பாடு அடையாளப்படுத்தல்.	167
4.7.1.	இந்தி மொழியிலிருந்து இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறையின் சொல் வகைப்பாட்டுக் குழுமம்.	168
4.7.2.	சொல்வகைப்பாடு அடையாளப்படுத்தும் பொறியின் செயல்பாட்டுத் தனிக்குறிப்பீடுகள்	171
4.7.2.1.	இடுவிரல் விடுவிரல் தனிக்குறிப்பீடுகள்	171
4.7.2.2.	சொல் வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம்	173
4.7.2.3.	செயல்பாட்டு வருணணைகள்	176
5	இயல் 5: தமிழ்த் தரவுத் தொகுதியில் தொடர் பகுத்தல்	177
5.1.	முன்னுரை	177
5.2.	பகுத்தல்	181
5.3.	ஆழமில்லாப் பகுத்தல்	181
5.3.1.	சொல்வகை அடையாளப்படுத்தல்	183
5.3.2.	தொடர் பகுத்தல்	183
5.3.3.	தொடர் உறவைக் கண்டுபிடித்தல்	184
5.4.	பெயர்த் தொடர் பகுத்தல்	184

5.5.	தமிழ்த் தரவுத் தொகுதியைத் தொடர் பகுத்தல்	185
5.5.1.	தமிழ்த் தொடரியல் அமைப்பு	185
5.5.1.1.	தொடர்கள்	187
5.5.1.1.1.	பெயர்த் தொடர்	187
5.5.1.1.2.	வினைத்தொடர்	187
5.5.1.2.	வாக்கிய வகைகள்	188
5.5.1.2.1	.அடிப்படை வாக்கிய வகைகள்	189
5.5.1.2.2.	வினா வாக்கியங்கள்	191
5.5.1.2.3.	செய்வினை, செயப்பாட்டு வினைவாக்கியங்கள்	195
5.5.1.2.4	கலப்பு வாக்கியம்	196
5.5.1.2.5	கூட்டு வாக்கியம்	200
5.5.2	முந்தைய கணினித் தொடரியல் ஆய்வுகள்	204
5.5.2.1	AU-KBC – RC-இன் தமிழுக்கான பெயர்த்தொடர் பகுப்பான்	207
5.5.2.2	RCILTS – இன் வானவில்	207
5.5.2.3	குமார சண்முகத்தின் தொடரியல் பகுப்பாய்வு	209
5.5.2.3.1	வாக்கியங்களின் இலக்கணப் பண்புக் கூறுகள்	209
5.5.2.3.2	இலக்கணப் பண்புகளும் அவற்றுடன் தொடர்புடைய அடையாளப்படுத்திகளும்	210
5.5.2.3.3	உருபனியல் பகுப்பாய்வு	211
5.5.3.4	சொல் வடிவமும் அதன் அடையாளங்களும்	212
5.5.3.5	சொல்வகை அடையாளப்படுத்தல்	212
5.5.3.6	இயல்பாக்கம்	213
5.5.3.7	இலக்கணச் சட்டகம்	214
5.6	இந்திய மொழியிலிருந்து இந்தியமொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர் பகுத்தல்	222

5.6.1.	இந்திய மொழியிலிருந்து இந்தியமொழி மொழிபெயர்ப்புத் திட்டத்திற்குத் தேர்ந்தெடுக்கப்பட்டுள்ள தொடர்கள்	223
5.6.1.1.	பெயர்த்தொடர்	224
5.6.1.2	வினைத்தொடர்	225
5.6.1.3	பெயரடைத் தொடர்	225
5.6.1.4	வினைத்தொடர்	225
5.6.1.5	இணைத்தொடர்	225
5.6.1.6.	வியப்புத்தொடர்	226
5.6.1.7	எதிர்மறைத் தொடர்	226
5.6.2	தொடர்பகுத்தலின் எடுத்துக்காட்டுகள்	226
6	இயல் -6: இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு இந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை	228
6.1	அறிமுகம்	228
6.1.1	நோக்கம்	228
6.1.2	வாய்ப்பு	230
6.1.3	சரியான தன்மை/பயன்படுத்துவோர் திருப்தி	230
6.1.4	ஒழுங்குமுறைச் சிறப்பு	230
6.1.5	நெறிமுறை	231
6.1.5.1	பகுதிகளாக அமைத்தல்	231
6.1.5.2	சக்தியின் தரமான அமைப்பு	232
6.1.5.3	தோல்விகளை நேர் செய்தல்	232
6.1.5.4	தெளிவு	232
6.1.5.5	டாஷ்போர்டு	233
6.2	முழுவிளக்கம்	233
6.2.1	உருபனியல் பகுப்பாய்வு	236
6.2.2	சொல்வகைப்பாடு அடையாளப்படுத்தி	237

6.2.3	தொடர் பகுப்பான்	237
6.2.4	வட்டாரச்சொல் குழுமி	237
6.2.5	மாற்றல் பகுதி	238
6.2.6	இருமொழிய அகராதி பொருத்தம்	239
6.2.7	வாக்கிய நிலை உருவாக்கம்	239
6.2.8	வட்டாரச்சொல் குழு பிரிப்பான்	240
6.2.9	உருபனியல் உருவாக்கம்	240
6.2.10	மதிப்பீடு செய்தல்	240
6.3	முடிவுரை	241
7	இயல் – 7 : முடிவுரை	243
	துணைநூற் பட்டியல்	246

இயல் 1

முன்னுரை

அன்றிலிருந்து இன்று வரை மொழிபெயர்ப்பு இன்றியமையாத மொழிச் செயல்பாடாக அமைகின்றது. பல மொழிகள் பேசப்படுகின்ற இந்தியா போன்ற நாடுகளில் மொழிபெயர்ப்பு ஒரு பொதுவான நிகழ்வாக நடந்துகொண்டிருக்கின்றது. ஒரு மொழியில் எழுதப்படுகின்ற மற்றும் பேசப்படுகின்ற எந்தத் தகவலும் செய்தியும் உடனுக்குடன் வேற்று மொழிகளில் மொழிபெயர்க்கப்படவேண்டிய கட்டாயம் நம்மிடையே இருந்து கொண்டிருக்கின்றது. பன்மொழிச் சூழலைப் பிரதிபலிக்கும் இந்தியா போன்ற நாடுகளில் மொழிபெயர்ப்பு ஒரு அன்றாடப் பணியாக நடந்து கொண்டிருக்கின்றது. மொழிபெயர்ப்பாளர்கள் எல்லாக் காலகட்டத்திலும் முக்கித்துவம் வாய்ந்தவர்களாக இருந்து வந்திருக்கின்றார்கள்; மொழிபெயர்ப்பும் எல்லாக் காலகட்டத்திலும் முக்கியத்துவம் வாய்ந்த செயல்பாடாக இருந்து வருகின்றது. நமது சிந்தனைகள் ஒலிக்குறியீடுகளாகவோ எழுத்துக் குறியீடுகளாகவோ மாற்றப்படுவதும் ஒரு வித பெயர்ப்பே. கணிப்பொறியின் வரவால் ஒரு மொழியிலுள்ள உரைகளை அல்லது பனுவல்களை மற்றொரு மொழிக்கு மாற்றும் பணி மற்றொரு பரிமாணத்தை நோக்கிப் பயணிக்கின்றது. மனிதச் செயல்பாடான மொழிபெயர்ப்பை கணிப்பொறி செயல்பாடக மாற்றும் கட்டாயத்திற்கு நாம் தள்ளப்பட்டுள்ளோம். இதன் காரணமாக இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் இன்றியமையாத செயல்பாடாகக் கருதப்பட்டு அதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

மொழிபெயர்ப்புச் செயல்பாட்டின் போது ஒரு மொழிபெயர்ப்பாளர் அறிந்திருக்கவேண்டிய தகவல்கள் யாவும் கணிபொறிக்குத் தந்தால்தான் இது சாத்தியமாகும். மொழிபெயர்ப்பவர் மூல மொழி அமைப்பு பற்றியும் இலக்கு மொழி

இலக்கண அமைப்பைப் பற்றியும் அறிந்திருக்கவேண்டும். எவ்வாறு மூல மொழியின் இலக்கண அமைப்பும் இலக்குமொழியின் இலக்கண அமைப்பும் வேறு படுகின்றன என அறிந்திருக்கவேண்டும். மூல மொழிச் சொற்களுக்கு இணையான இலக்கு மொழிச் சொற்கள் என்ன? மூல மொழித் தொடர்களுக்கு இணையான இலக்கு மொழித் தொடர்கள் என்ன? மூல மொழி வாக்கியங்களுக்கு இணையான இலக்குமொழி வாக்கியங்கள் என்ன என்பன மொழிபெயர்ப்பாளர் அறிந்திருக்கவேண்டும். மூலமொழியை இலக்கு மொழியாக மாற்றப் பணித்துள்ள கணிப்பொறியும் இவ்வறிவைப் பெற்றிருக்கவேண்டும். மூல மொழியின் மற்றும் இலக்குமொழியின் உருபனியல் அமைப்பு, சொல்லியல் அமைப்பு, தொடரியல் அமைப்பு, பொருண்மையியல் அமைப்பு என்பன பற்றியும் எவ்வாறு இவ்வமைப்புகள் மூல மொழியிலும் இலக்கு மொழியிலும் வேறுபடுகின்றன என்பது பற்றியும் கணிப்பொறிக்கு அறிவுறுத்த வேண்டும். மொழி உரைகள் வெறும் எழுத்துக்களின் கோர்வையல்ல, எழுத்துக்கள் உருபன்களாகவும் உருபன்கள் சொற்களாகவும் சொற்கள் தொடர்களாகவும் தொடர்கள் வாகியங்களாகவும் பகுப்பாய்வு செய்யப்பட்டுதான் எழுதுக் கோர்வைகள் பொருள் பெறுகின்றன என்ற நமது அறிவைக் கணிப்பொறிக்குத் தரவேண்டும்.

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்ப்பு ஒழுங்குமுறை (Indian Language to Indian Language Machine Translation (IL-ILMT System)) மற்றும் இந்திய மொழிகளுக்கான தரவுத்தொகுதி உருவாக்க முயற்சி (Indian languages corpora Initiative (ILCI) என்ற மைய அரசின் கருத்துப்பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் நிதி நல்கையில் எனது தலைமையின் கீழ் தமிழ்ப்பல்கலைகழக்த்தில் நடைபெற்றுவந்த ஆய்வுத்திட்டங்கள் மூலம் நான் பெற்ற அனுபவத்தின் வெளிப்பாடே இவ்வேடாகும்.

மூலமொழி அமைப்பு பற்றிய அறிவையும் இலக்கு மொழி அமைப்பு பற்றிய அறிவையும் அவைகள் எவ்வாறு வேறு படுகின்றன என்ற அறிவையும் அதன் அடிப்படையில் மூல மொழி உரைகளை இலக்கு மொழி உரைகளாக மாற்றும்

விதிகளையும் கணிப்பொறிக்குத் தந்து கணிப்பொறியை மொழிமாற்றம் செய்யப் பணித்தால் கணிப்பொறி அப்பணியைச் செவ்வனே செய்யும் என்பது எதிர்பார்ப்பாகும். ஆனால் எதிர்ப்பார்ப்புக்கும் சாத்தியத்திற்கும் விளைவிற்கும் இடையில் பெரிய இடைவெளிகள் உள்ளன. கணிப்பொறிக்கு விதிகளைச் சொல்லித் தருவதைவிட விதிகளை தானாகவே உணரச் செய்யும் செயல்பாடு எளிதானதும் சாத்தியமானதாக அமைகின்றது. இத்தகைய சூழலில் புள்ளியியல் செயல்பாடு முக்கியத்துவம் பெறுகின்றது. கணிப்பொறிக்கு எவ்வாறு செய்யவேண்டும் என்ற பயிற்சியை அளித்துவிட்டால் அது தானே நாம் விரும்பும் செயல்பாட்டைத் தவறின்றிச் செய்யத் தொடங்கும். தேவையான அளவு சொல் எண்ணிக்கையுள்ள தமிழ்த் தரவுத் தொகுதிகளைச் சேகரித்து சொல் வகை அடையாளப் படுத்தலையும் தொடர்ப் பிரித்து அடையாளப்படுத்தலையும் செய்து கணிப்பொறிக்குத் தந்தால் அது அச்செயற்பாடுகளைத் தானாகவே கற்றுக்கொண்டு புதிய தரவுத்தொகுதிகளைச் சொல்வகைக்கு அடையாளப்படுத்துவதையும் தொடர்களாகப் பிரிப்பதையும் தானாகவே செய்யும். இதன் அடிப்படையில் தான் தரவுத் தொகுதிகளைச் சேகரித்து சொல்வகை அடையாளப்படுத்துவதையும் தொடர்பிரித்து அடையாளப்படுத்துவதையும் மனித முயற்சியால் செய்து அதைக் கணிப்பொறிக்குக் கற்பித்து அந்த அறிவைப் பயன்படுத்தி கணிப்பொறி புதிய தரவுத் தொகுதிகளை சொல்வகைக்கு அடையாளப்படுத்தவும் தொடர்பிரித்து அடையாளப்படுத்தவும் செய்யும்படிக்கு பயிலுவிப்பது முக்கியச் செயல்பாடாக அமைகின்றது. சொல்வகைப்பாடு அடையாளப்படுத்தலையும் தொடர்பிரித்து அடையாளப்படுத்தலையும் மனித முயற்சியால் செய்து அவ்வடையாளப்படுத்தல்களைக் கணிப்பொறிக்குக் கற்றுத்தந்தால் கணிப்பொறி இவ்வடையாளப்படுத்தும் செயல்பாடுகளைச் செவ்வனே செய்யவியலும் என்பது இயந்திரம் கற்றல் நுட்பாமாகும்.

தமிழுக்கான தரவுத் தொகுதிகள் பல உருவாக்கப்பட்டாலும் மொழியியல் கூறுகளுக்கு அடையாளப்படுத்தப்பட்ட முறையான தரவுத்தொகுதி இதுவரை தமிழுக்கு உருவாக்கப்படவில்லை. தரவுத் தொகுதி பற்றிய விரிவான விளக்கமும்

இதுவரை தமிழில் உருவாக்கப்பட்டுள்ள தரவுத் தொகுதிகள் பற்றி இங்கு தரப்பட்டுள்ள தகவல்களும் இதற்கு வழிவகுக்கும்.

தரவுத் தொகுதியைச் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்துவதும் தொடர் உறுப்புகளாகப் பகுப்பதும் பல இயற்கைமொழி ஆய்வுச் செயல்பாடுகளுக்கு இன்றியமையாத செயல்பாடுகளாகும். சொல்வகைப்பாட்டிற்கும் தொடர் உறுப்புக்கும் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பல கணினி மொழியியல் தொடர்பான பயன்பாடுகளுக்கு உறுதுணையாக அமையும் என்ற அடிப்படையில் இவ்வேடு முக்கியமானதாக அமைகின்றது. மேலும் இயந்திர மொழிபெயர்ப்பு பற்றியும் தமிழில் இயந்திர மொழிபெயர்ப்பு பற்றியும் இங்கு தரப்பட்டுள்ள தகவல்கள் பிற்கால ஆய்வுகளுக்கு அடித்தளமாக அமையும்.

இயல் 2

இயந்திர மொழிப்பெயர்ப்பு

2.1 அறிமுகம்

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப் பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களைப் பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. வெறும் கணிக்கும் பொருளாக இருந்த கணிப்பொறி தரவுகள் சேமித்து வைக்கும் கருவியாகவும் தகவல் தொடர்பு கருவியாகவும் தரவுளைத் தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகிறது. மொழிச் செயல்பாட்டிற்கும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவியல் உள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கானச் சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத்தட்டு, பதிவுநாடா போன்றவற்றில்) சேகரித்து வைத்துத் தேவையாபோது பயன்படுத்திக் கொள்ள இயலும்.

மொழிக்கான தரவுகள் பல வகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவோ, வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். இவைகள் உள்ளடக்கிய ஒரு உரையாகவும் (text) அமையலாம். வெவ்வேறு வகைப்பட்ட உரைகளிலிருந்து தேர்ந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு விரிதரவவு (Corpus) என்று அழைக்கப்படும். தமிழுக்கான விரிதரவைச் சேகரித்து வைப்பது மிக இன்றியமையாதப் பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய உரைகள் வட்டார வழக்கு, கிளைமொழி வழக்கு, சமுதாய வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய உரைகள் எல்லாம் ஒருசேர விரிதரவாக அமையும் இத்தகைய விரிதரவை இருவகையாகப் பிரிக்கலாம்.

1. எழுத்து வடிவில் அமைந்த தரவுத்தொகுதி (Written corpus)
2. பேச்சு வடிவில் அமைந்த தரவுத்தொகுதி (Speech corpus)

இத்தகைய தரவுத்தொகுதிகள் மொழியாய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவி புரிகின்றன.

தரவுகள் கணிப்பொறியால் ஆயப்பட்டுத் தரப்படும். மொழித் தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வனதட்டு, குறுந்தட்டு, பதிவு நாடா போன்றவற்றில்) சேமித்து வைத்து தேவைப்படும் போது பயன்படுத்திக் கொள்ளலாம். பல்வேறு விதமான தகவல்ககள் இணையத்தளத்தில் நிரம்பிக்

கிடக்கின்றன். தமிழ்மொழியின் கட்டமைப்பைக் கணிபபொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ் மொழியைப் பயன்படுத்தி தகவல்களைப் பெற உதவும். தகவல் தேடும் பொறி (Search engine) தகவல்களைப் பெற உதவும் முக்கிய கருவியாகும். தமிழுக்கு இத்தைகைய தகவல் தேடும் பொறி பெருமளவில் வெற்றியும் அடைந்துள்ளது.

மொழி உரைகளைப் படிநிலை (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழி அலகுகளாகப் பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்து கொளள கணிப்பொறையைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக் கொள்வோம்.

எ.கா.

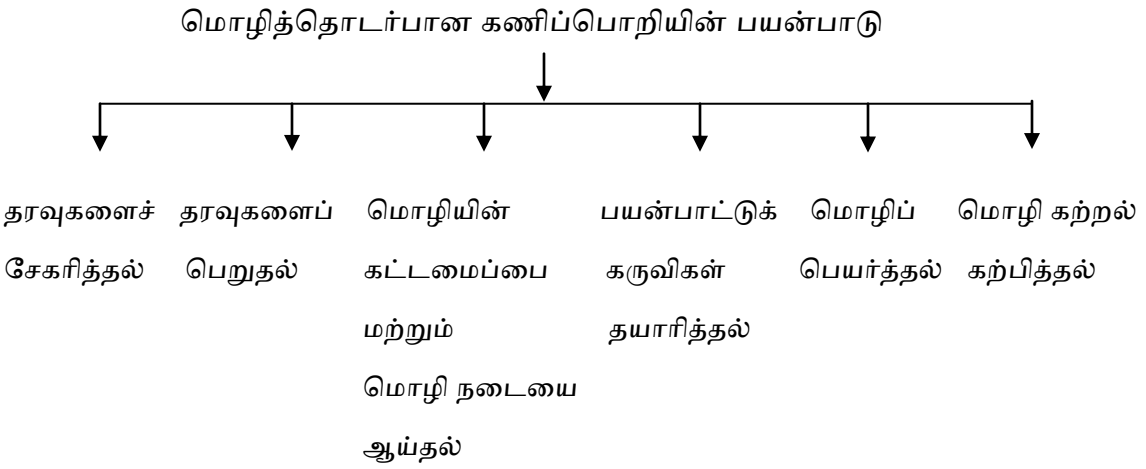
பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்.

பள்ளி மாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும், *பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத்தொடர் என்றும், *பள்ளி, மாணவன், மரம், பலாப்பழம், ரகுமான்* என்பன பெயர்கள் என்றும், பறி என்பது வினை என்றும், *பெரிய* என்பது பெயரடை என்றும், கவனமாக என்பது வினையடை என்றும், இதிலிருந்து என்பது பின்னுருபு என்றும், ஐ-என்பது வேற்றுமையுருபு என்றும் *–த்-* என்பது இறந்தகால இடைநிலை என்றும் *-ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிப்பொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுப்பாய்வுச் செய்து தந்து விடும்.

ஒரு கால கட்டத்தில் மொழியலார் கள ஆய்வு செய்து தகவலாளிகளின் உதவயால் ஆயப்படாத மொழிகளின் தரவுகளைச் சேகரித்து அதன் மீத சில மொழியியல் கொள்கைகளையும் கோட்பாடுகளையும் பயன்படுத்தி சில கருதுகோள்களின் அடிப்படையிலும் உற்றுநோக்கல் மூலமும் அம்மொழிகளின் ஒலியன், உருபன், தொடர், பொருண்மை அமைப்புகளைக் கண்டுபிடிக்க முயலுவர். இது பண்புசார் மொழியியல் (Qualitative linguistics) எனப்படும். இப்போது கணிப்பொறியின் வரவால் அத்தகைய நிலை மாறி மொழிகளின் தரவுத்தொகுதிகள் உருவாக்கப்பட்டு அதன் மீது புள்ளியியல் நெறிமுறைகளின் அடிப்படையில் மேற்சொன்ன மொழி அமைப்புகள் கண்டுப்பிடிக்கப்படும். இதை அளவுசார் மொழியியல் (Quantitive Linguistics) என்பர். தற்போது மொழி ஆய்வுகள் எல்லாம் கணினியைப் பயன்படுத்தி தரவுத்தொகுதிகள் மீது செய்யப்படும் ஆய்வுகளாக மாறியள்ளன.

மொழி உரையை இடுவிரலாகத் (input) தந்து தக்க வழியமைப்புகளும் (Programs)கட்டளைகளும் தந்தால் மொழியின் உரைகளை வாக்கியங்களாகவும் வாக்கியங்களை தொடர்களாகவும், தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக் கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்கானத் தொடரியல் ஆய்வி (Syntactic analyzer), உருபனியல் ஆய்வி (Morphological analyzer), ஒலியனியல் ஆய்வி (Phonological analyzer), பொருண்மையியல் ஆய்வி (Sematic analyzer) என்ற ஆய்விகளைத் தயாரித்து பயன்பாட்டில் பயன்படுத்துகின்றன.

மொழித்தொடர்பானக் கணிப்பொறியின் பயன்பாட்டை ஆறு முக்கிய அம்சங்களாகப் பகுக்கலாம்.



இவ்வமைப்பின் படி மொழி ஆய்வில் கணிப்பொறியின் பங்கு வியக்கும் அளவில் இருக்கிறது என்பதை நம்மால் உணரமுடிகிறது. இக்காலத்தில் எந்த செயலைச் செய்வதற்கும் கணினி இல்லாமல் செய்ய முடியாது. அந்த அளவிற்கு கணினியின் பயன்பாடு பேருதவி புரிந்து வருகின்றது. மொழியாயவில் கணிப்பொறி முதல் கருவியாகக் கொண்டு விரைவில் ஒரு ஆய்வை குறித்த நேரத்தில், குறிப்பிட்ட காலத்தில் செய்து முடிப்பதற்கு கணினியின் பங்கு முக்கியமானதாகக் கருதப்படுகிறது.

இயந்திர மொழி பெயர்ப்பு இன்றைய கணினி யுகத்தில் மிக முக்கியமானதாகக் கருதப்படுகிறது. பல நூற்றாண்டுகளாக மனிதனால் செய்யப்பட்ட மொழி பெயர்ப்பு செயல்பாட்டைக் கணினியால் எவ்வாறு எளிதாகவும் விரைவாகவும் செய்யலாம் என்ற முயற்சிகள் 1950-களிலேயே தொங்கப்பட்டு பின்னர் கைவிடப்பட்டு மீண்டும் புதிய திசையில் வளர்ந்து கொண்டிருக்கிறது. கணிப்பொறியின் தொழில் நுட்பத்தில் ஏற்பட்ட முன்னேற்றங்கள் இயந்திர மொழி பெயர்ப்பிலும் முன்னேற்றத்தை ஏற்படுத்தியது, பலவிதமான மாதிரிகள் அடிப்படையில் இயந்திர மொழி பெயர்ப்பு செய்யப்பட்டு பரிசோதனைக்கு உள்ளாக்கப்பட்டது. நேரடியான சொல்லிலிருந்து சொல்

மொழிபெயர்ப்பு மாற்றம், தொடரிலிருந்து தொடர் மொழி பெயர்ப்பு மாற்றம், பொருண்மையிலிருந்து பொருள் மொழி பெயர்ப்பு மாற்றம் எனப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

கணிப்பொறியை மட்டும் வைத்துக்கொண்டு மொழிப்பெயர்ப்பு செய்ய இயலாது என்று புரிந்துகொள்ளப்பட்டதன் விளைவாகக் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு, மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு, முற்றிலும் தானியங்கி மொழி பெயர்ப்பு எனப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

பலவிதமான இலக்கண வடிவமைப்புகள் (Grammatical Formalisms) மொழி பெயர்ப்பிற்குப் பயன்படுத்தப்பட்டன. எடுத்துக்காட்டாக முற்றுநிலைத் தானியங்கி, சூழல் கட்டிலா இலக்கணம் (Context Freee Grammar) சூழல் கட்டுண்ட இலக்கணம் (Context Sensitive Grammar), ஆக்கமுறை மாற்றிலக்கணம் (Generative Transfomational Grammar), கிளை சேர்ப்பு இலக்கணம் (Tree Adjoining Grammar), தொடரமைப்பு இலக்கணம் (Phrase Structure Grammar), தலைமை இயக்கத் தொடரமைப்பு இலக்கணம் (Head Driven Pharse Structurs Grammar (HPSG)), பாணினி இலக்கணம் (Panini Grammar), வேற்றுமை இலக்கணம் (Case Grammar), சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), பொதுமைப்படுத்தப்பட்ட தொடரமைப்பு இலக்கணம் (Generalized Phrase Strcture Grammar (GPSG) எனப்பல இலக்கண வடிவமைப்புகள் சொல்லமைப்பு ஆய்விற்காகவும் தொடரமைப்பு ஆய்விற்காகவும் பயன்படுத்தப்பட்டன.

தமிழ் போன்ற உருபனியல் அடிப்படையில் வளமுள்ள மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கப்பட வேண்டுமானால் உருபனியல் பகுப்பாய்வு செம்மையாகச் செய்யப்பட்ட வேண்டும் எனக் கருதி உருபனியல் பகுப்பாய்விகள் (Morphological Analyzers) உருவாக்கப்பட்டன.

உருபனியல் பகுப்பாய்வி இயந்திர மொழி பெயர்ப்பின் முக்கியமான ஒரு பாகமாகக் கருதப்பட்டு அதை உருவாக்கப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

ஆங்கிலம் போன்ற மொழிகளுக்கு உருபனியல் வளத்தைக் காட்டிலும் தொடரியல் வளம் கூடுதலாக இருந்தன காரணமாகத் தொடரியல் பகுப்பாய்விற்கு முக்கியம் தரப்பட்டுத் தொடரியல் பகுத்துக் குறிப்பான்கள் (Syntactic Parsers) உருவாக்கப்பட்டன.

உருபனியல் பகுப்பாய்வும் தொடரியல் பகுப்பாய்வும் மூலமொழியில் செய்யப்பட்டு அதில் வரும் வெளியீடு இலக்கு மொழியில் மாற்றப்படும். இங்கு சிக்கலான செயல்பாடு என்னவென்றால் உருபனியல் பகுப்பாய்வின் வெளியீடு ஒன்றிற்கு மேற்பட்ட அர்த்தத்தைத் தரும் உருபனியல் பகுத்துக் குறித்தல்களையும் தொடரியல் பகுப்பாய்வின் வெளியீடு ஒன்றிற்கு மேற்பட்ட அர்த்தத்தைத் தரும் தொடரியல் பகுத்துக் குறித்தல்களைத் தருவதும் ஆகும். அதாவது மூலமொழியில் இருக்கும் ஒரு சொல்லுக்கு ஒன்றிற்கும் மேற்பட்ட உருபனியல் பகுத்துக் குறித்தல்களும் மூலமொழி வாக்கியத்திற்கு ஒன்றிற்கும் மேற்பட்ட தொடரியல் பகுத்துக் குறித்தல்களும் வருவது சாத்தியமாகும். ஒரு மூலமொழி வாக்கியத்திலிருந்து கிடைக்கும் ஒன்றிற்கு மேற்பட்ட பகுத்துக் குறித்தல்களில் ஒன்றுதான் சரியானது. அந்தச் சரியான ஒன்றுதான் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மாற்றப்படவேண்டும். எனவே பொருள் மயக்கம் நீக்கப்பட வேண்டிய கட்டாயம் ஏற்படுகிறது. இதைச்சொல் மயக்க நீக்கம் (Words Sense disambiguation) என்பர். சொல் மயக்க நீக்கம் சூழல் அடிப்படையில் செய்யப்பட வேண்டியதாகும். இதற்காகப் பல புள்ளியியல் மாதிரிகள் இருக்கின்றன. இத்தகைய புள்ளியியல் மாதிரிகளைப் பயன்படுத்தி சொல் மயக்கம் நீக்கிச் சரியான தொடரியல் பகுத்துக் குறித்தலைப் பெற்று மூலமொழி இலக்கு மொழிக்கு மாற்றம் செய்யப்பட வேண்டும். மேற்கூறிய இயந்திர மொழிபெயர்ப்புச்

சிக்கல்களையும் இயந்திர மொழி பெயர்ப்பின் எல்லைகளையும் கருத்தில் கொண்டு மொழி பெயர்ப்பிற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன.

மொழிபெயர்ப்பின் வெளிப்பாடு தெளிவாக இருக்க வேண்டும். சொற்பெயர்ப்பு மூல நூலின் சொற்களுக்கு இணையானதாகவும் எளிமையானதாகவும் சொற்கட்டு உடையதாகவும் இருக்க வேண்டும். மூல நூலில் உள்ளது போலவே கருத்துப் புலப்பாட்டுத் திறன் சிறப்பாக அமையவேண்டும். வாய்பாடு (Formula), சமன்பாடு (Equation), குறியீடு (symbol), சுருக்கக்குறியீடு அகியவற்றை அப்படியே பயன்படுத்தலாம். plan, scheme, project, programme என்பவற்றைத் 'திட்டம்' என்றே மொழிபெயக்கின்றனர். இச்சொற்களிடையே நுண்ணிய வேறுபாடுகள் உள்ளன. சொற்களுக்கிடையே உள்ள நுண்ணிய வேறுபாடுகளை அறிந்து மொழிபெயர்க்க வேண்டும். மூல மொழியிலிருந்து கருத்துக்களைப் பெயர்க்கும் போது சொற்பெருக்கம் நிகழ்கிறது. கடன் சொற்கள், கலப்புச் சொற்கள், கூட்டுச் சொற்கள், பழஞ்சொற்களுக்குப் புதுப் பொருள் எனப் பலவிதமாக இது அமைகின்றது. இதனால் மொழியமைப்பு விரிவடைகின்றது. புதுப் புது சொல்லகுகள் கிடைக்கின்றன. மொழியில் இலக்கண விரிவாக்கமும் நிகழ்கின்றது. மூல மொழியிலிருந்து குறிக்கோள் மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது குறிக்கோள் மொழியமைப்புக்கேற்ப மொழிபெயர்ப்பு அமைய வேண்டும். மொழிபெயர்ப்பு ஒரு மொழியிலிருக்கும் அறிவியல் கருத்துக்களை வேற்றுமொழிக்குப் பரிமாற்றம் செய்வதற்கு உதவுகின்றது. மொழிபெயர்ப்பு அறிவு வளர்ச்சிக்கும் மொழி வளர்ச்சிக்கும் மிகவும் உதவுகின்றது. கலைச் சொல்லாக்கத்தில் மொழிபெயர்ப்புதான் பெரும் பங்கு வகிக்கின்றது. தமிழில் கலைச்சொற்கள் எல்லாம் ஆங்கில மொழிபெயர்ப்பாகும்.

எ.கா.

Telephone :தொலைப்பேசி

Television :தொலைக்காட்சி

University :பல்கலைக்கழகம்

Photosynthesis :ஒளிச்சேர்க்கை

Electron :மின்னணு

இம்மொழிபெயர்ப்பு முற்காலத்தில் மனித முயற்சியால் மேற்கொள்ளப்பட்டு வந்தது. இதனால் காலவிரையமும் எளிதில் மொழிபெயர்க்க இயலாத நிலையும் ஏற்பட்டது. தற்காலத்தில் கணினியின் வரவால் மொழிபெயர்ப்புப் பணி முன்னேற்றம் அடைந்து கணினியைப் பயன்படுத்தி மொழிப்பெயர்ப்பு செய்யும் முறை மேற்கொள்ளப்பட்டு நடைப்பெற்று வருகின்றது.

2.2 கணிப்பொறி வழி மொழி பெயர்ப்பின் தேவை

ஒரு மொழியில் உள்ள தகவலை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதையே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பில் தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியின் இயல்புத் தன்மை மாறாது பாதுகாத்தல் வேண்டும். தகவல் எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Language) என்கிறோம். தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகின்றதோ அந்த மொழியை இலக்குமொழி (Target Language) என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியால் மட்டும் மேற்கொள்வது என்பது கடினமானப் பணியாகும்.

எனவே மொழிபெயர்ப்புப் பணிகளைக் கணிப்பொறி வழி மேற்கொள்வதற்கான கட்டாயம் ஏற்படுள்ளது. கணிப்பொறி வழி மொழிபெயர்ப்பின் தேவைகளும் கட்டாயங்களும் கீழேப் பட்டியலிடப்பட்டுள்ளன.

- உலகளவில் தகவல்கள் அன்றாடம் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழிபெயர்த்தல் அவசியமாகிறது.

- கணிப்பொறி வழி மொழிபெயர்ப்பு மனித உழைப்பைக் குறைக்கிறது.
- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்புப் பணிகளைச் செய்ய இயலுகிறது. இதனால் காலவிரயம் தவிர்க்கப்படுகிறது.
- தற்போதைய மொழி பெயர்ப்புப் பணிகளுக்கு ஏற்ற வகையில் மொழிப் பெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மொழிபெயர்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழியிலும் இலக்கு மொழியிலும் புலமை உள்ளவர்களிடம் கருத்துக்களைப் பெற்று அதை வழியமைப்பு மொழயில் எழுதி மொழிபெயர்ப்பு செய்யலாம்.

2.3 கணிப்பொறி வழி மொழிபெயர்ப்பின் வகைகள்

மொழிபெயர்ப்பில் கணிப்பொறி எவ்விதமாகப் பயன்படுத்தப்படுகின்றது என்பதன் அடிப்படையில் மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்:

1. கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு
   (CAMT – Computer Assisted Human Translation)
2. மனித உதவியின் கணிப்பொறி மொழிபெயர்ப்பு
   (HAMT-Human Assisted Machine Translation)
3. முற்றிலும் தானியங்கி மொழிபெயர்ப்பு
   (Automatic Machine Translation)

2.3.1 கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனிதமொழி பெயர்ப்பில் மனித மொழிப் பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகின்றது. அவர்கள் மொழிப்பெயர்ப்புப் பணியில் ஈடுபடும் பொழுது ஒரு மொழியில் உள்ள

சொற்களுக்கு நிகரான பிறமொழிச் சொற்களைத் தெரிந்துகொள்ளவும் அச்சொற்களின் பொருள் பயன்பாடு ஆகியவற்றை எளிமையாகத் தெரிந்துகொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாகத் தயார் நிலையில் பெறுவதற்கும் கணிப்பொறி உதவுகிறது. இவ்வகை மொழிபெயர்ப்பில் மொழிப்பெயர்பாளர்கள் தான் முதல் நிலையில் செயலாற்றுவர்; அவர்களுக்கு உதவும் நிலையில் கணிப்பொறி செயலாற்றும்.

2.3.2 மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிப்பெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிப்பெயர்ப்புப் பணிகளைச் செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ள நகலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிப்பெயர்ப்பு வகைகளில் மனித உதவியுடன் செயல்படும் இயந்திர மொழி பெயர்ப்பு தான் நடைமுறையில் பரவலாக உள்ளது.

2.3.3 முற்றிலும் தானியங்கி மொழிபெயர்ப்பு

முற்றிலும் இயந்திர மொழிபெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல இதற்கு எண்ணற்ற மொழியில் செயல்பாடுகளும் மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட்ட வேண்டி உள்ளது. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பில் முற்றிலும் தானியங்கப்படுத்தும் முயற்சி இலக்காகக் கொள்ளப்பட்டுள்ளது.

2.4 கணிப்பொறி வழி மொழிப்பெயர்ப்பின் சில முயற்சிகள்

கணிப்பொறி வழி மொழி பெயர்ப்பினைத் தானியங்கி மொழிப்பெயர்ப்பு (Automatic Translation) என்றும் இயந்திர மொழிபெயர்ப்பு (Machine Translation) என்றும் கணிப்பொறி உதவியுடன் நடைபெறும் மொழிபெயர்ப்பு என்றும் கூறலாம். கணிப்பொறியின் வேகம், நினைவகத்திறன், பதிவு செய்துள்ள செய்திகளைத் தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளப்பெறக்கூடிய வசதி (Retrievel System) போன்றவைக் கணிப்பொறி வழி மொழி பெயர்ப்புப் பணியை ஊக்குவிக்க உதவின.

கணிப்பொறி வழி மொழிபெயர்ப்புப் பணி 1950-ஆம் ஆண்டு வாரன் வீவர் (Warren Waiver) என்ற அறிஞரால் தொடங்கப்பட்டது. இதைத் தொடர்ந்து பூபார் ஹில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணியைத் தொடங்கினார். 1954-ஆம் ஆண்டு ஜார்ஜ் டவுன் பல்கலைக்கழகம் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியைத் தொடங்கியது. இதைத் தொடர்ந்து 1956-ஆம் ஆண்டு மாஸ்கோவில் பிரெஞ்சு மொழியிலிருந்து ரஷ்ய மொழிக்குக் கணிப்பொறி வழி மொழிப்பெயர்ப்பு செய்யும் பணி நடைபெற்றது. இப்பணியை மேல் சங் (Mel Chunk) என்ற அறிஞர் மேற்கொண்டார்.

1973-ஆம் ஆண்டு அமெரிக்காவில் உள்ள Logos Incoration Limited என்ற நிறுவனம் ஒரு மென்பொருளை வடிவமைத்து. இம்மென் பொருளைக் கொண்டு ஆங்கிலத்தில் அமைந்துள்ள செய்திகளை வியட்நாம் மொழியில் மொழிப்பெயர்ப்பு செய்தது. ஏறக்குறைய இதே காலக்கட்டத்தில் Tertils என்ற நிறுவனம் TUTUS என்ற மென்பொருளை வடிவமைத்தது. இம்மென் பொருளைக் கொண்டு ஆங்கிலம், ஸ்பானிஷ், பிரெஞ்சு மற்றும் ஜெர்மன் ஆகிய மொழியில் உள்ள செய்திகள் ஒன்றிலிருந்து மற்றவைகளுக்கு மொழிமாற்றம் செய்யப்பட்டது. இதைத் தொடர்ந்து CUTD-Chinese University Translator) என்ற மென்பொருள்

உருவாக்கப்பட்டது. இதைக்கொண்டு சீனமொழியில் உள்ள செய்தித்தாள்கள் ஆங்கில மொழிக்கு மொழிமாற்றம் செய்யப்பட்டன.

1974-ஆம் ஆண்டு SYSTRAN என்ற மென்பொருள் உருவாக்கப்பட்டது. SYSTRAN என்பதன் விரிவு System Translation என்பதாகும். இம்மென்பொருள் ரஷ்ய மொழியிலிருந்து பிரஞ்சு மொழிக்கு மொழி மாற்றம் செய்யப்பயன்பட்டது. இதற்கானத் திட்டத்தை P.Toma என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன்முதலில் வெளிப்யிடப்பட்ட கணிப்பொறி வழி மொழிபெயர்ப்புக்கான மென்பொருள் இதுவேயாகும்.

1975-ஆம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென் பொருள் ரஷ்ய மொழியில் மாற்றம் செய்யப்பட்டது. இதே காலக்கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு வழியமைக்க உருவாக்கப்பட்டது. இது போன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனைக்கட்ட முயற்சியாலும் வணிக நோக்கத்தாலும் கணிப்பொறி வழி மொழிபெயர்க்க மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைபெற்று வருகின்றன.

1985-ல் தமிழ்ப்பல்கலைக் கழகத்தில் TUMTS (Tamil University Mechine Translation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. ரஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழி பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப் பல்கலைக்கழகக் கணிப்பொறி அறிவியல் துறை, மொழியியல் துறை, மொழிப்பெயர்ப்புத்துறை ஆகியவைகளின் கூட்டு முயற்சியில் நடைமுறைப்படுத்தப்பட்டது. இதற்குப்பின்னால் கான்பூரில் உள்ள இந்திய தொழில்நுட்ப நிறுவனம் Indian Institute of Technology இந்தியிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து இந்திக்கும் மொழிபெயர்ப்பு செய்வதற்கும் இந்திய மொழிகளை இந்திக்கும் இந்தியை பிற இந்திய மொழிகளுக்கும்

மொழிபெயர்ப்பதற்கான ஆரம்ப கட்ட வேலைகளை மேற்கொண்டு செயலாற்றியது. அனுசாரகா (*anusaraka*) என்ற இயந்திர மொழிபெயர்ப்பு மென்பொருள் இதன் விளைவாகும்.

2.5 கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி மாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்ததைய கணிப்பொறி வழி மொழி பெயர்ப்புப் பணிகளில் எல்லைகள் இரண்டாக வரையறுத்து உள்ளன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக்கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்தக்கூடிய உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழி நடை இலக்கியமொழி, பேச்சுமொழி என இருவகையாகப் பிரிக்கப்படுகிறது. பேச்சுமொழியைக் கணிப் பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிப்பெயர்ப்புப் பணியில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக்கொள்ளப்படுகின்றன.

2. செய்யுள், கவிதை போன்றவைகளில் பொருளுக்குத் தரும் முக்கியதுவத்தைவிட கருத்தைப் புலப்படுத்தும் உத்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகின்றன. இவைகள் மனித உளவியலைப் பிரதிப்பலிப்பன. இவ்வுளவியல் கூறுகளைக் கணிப்பொறி வழி வெளிப்படுத்துவது என்பது இயலாத ஒன்றாகும். இதேபோல் இலக்கியங்களிலும் பொருளைவிட உளவியல் கூறுகளுக்கு தான் முக்கியத்துவம் தரப்படுகின்றது. எனவே, செய்யுள்கள், கவிதைகள்,

பாடல்கள் மற்றும் இலக்கியச் செய்திகள் ஆகியவைகளைக் கணிப்பொறி வழி மொழிபெயர்ப்பதற்குத் தொடக்கநிலையில் பயிற்சி எடுத்துக்கொள்ளப்படவில்லை. அறிவியல் தொழில்நுட்ப சார்ந்த படைப்புகளில் பொருள்களுக்கே முக்கியத்துவம் தரப்படுகிறது. இங்கு கோபம், வீரம், தாபம், மகிழ்ச்சி போன்ற உணர்வுகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணிகளுக்குப் பொதுவாக அறிவியல் மற்றும் தொழில்நுட்பத்தைச் சார்ந்த படைப்புகளே எடுத்துக்கொள்ளப்படுகின்றன.

இதேப் போன்று மொழிகளின் தன்மையைப் பொறுத்துக் கணிப்பொறி மொழிபெயர்ப்பினை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1. இருதொடர்புடைய மொழிகளுக்கிடையே அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Interlingual Method of Machine Translation): இலக்கண அமைப்பிலும் பொருள் அமைப்பிலும் தொடர்பும் நெருக்கமும் உடைய இருமொழிகளுக்கிடையே செய்யப்படும் கணிப்பொறி மொழிப்பெயர்ப்பு இவ்வகையைச் சார்ந்தாகும்.

2. இடைநிலை மொழி அடிப்படையில் அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Intermediate Language Based Machine Translation) இலக்கணம் மற்றும் மொழியமைப்பில் முற்றிலும் வேறுபட்ட இருமொழிகளுக்கிடையே கணிப்பொறி வழி மொழிபெயர்ப்பை நடைமுறைப்படுத்த இம்முறை உதவுகிறது. இம்முறையில் மூலமொழியில் உள்ள சொற்றொடர்கள் இடைநிலை மொழி வழி இலக்கு மொழிக்கு மாற்றம் செய்யப்படுகின்றது.

2.6 கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழிபெயர்ப்புப் பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற பொழுது கணிப்பொறியின் சிறப்புத் தன்மை, துல்லியம், வேகம், நினைவகக் கொள்திறன், தகவல் தளம் உருவாக்கும் வசதிகள் போன்றவை மொழிபெயர்ப்புத் திறனை நிர்ணயம் செய்யும். ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வழியமைக்கப்படுகிறது. மனித மொழிபெயர்ப்புப் பணிக்குக் கூட ஒருவருக்கு இருமொழியிலும் திறமை இருக்கவேண்டும் என்ற அவசியம் இல்லை.

ஒரு மொழியின் வளம், தன்மை ஆகியவற்றைக் கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையான பணி அல்ல. எனவே, கணிப்பொறி வழி மொழி பெயர்ப்புத் திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப்பட்டிருப்பினும் இன்னும் முழுமைபெறவில்லை. மேலும் கணிப்பொறி அறிவியலில் தொழில் நுட்ப வளர்ச்சி கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிக்கு மகிவும் பயனுள்ளதாக அமைகின்றது. இத்தகைய பணிகளின் நிறை குறைகளைக் கீழேக் காணலாம்.

மொழிக்கு இடைடே உள்ள சொல் வளம் இலக்கணம் கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிபெயர்ப்புப் பணியின் சிக்கல்கள் குறையும்.

உருப்பனியல் அமைப்பை பொறுத்தவரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

அ. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து.

ஆ. ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தாலும் அதற்கு மேலே இருந்தாலும் அவைகளைத் துல்லியமாகப் பிரிக்க இயலும் என்பதைப் பொறுத்து.

ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இந்தால் அந்த மொழியைத் தனிநிலை உருபன்கட்டு மொழி (isolating Language) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டுமொழி (Poly Synthetic) என்கிறோம். சிபேரியன் மொழி பன்நிலை உருபன் கட்டுமொழிக்குச் சிறந்த உதாரணமாகும். வியட்நாமிஸ் மற்றும் கேண்டுனஸ் ஆகிய மொழிகள் தனிநிலை உருபன்கட்டு (isolating) மொழிக்குச் சிறந்த உதாணரங்களகும். எனவே இத்தகைய மொழிக்குச் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்ற பொழுது மொழியின் உருபனியல் பற்றிய விரிவான தகவல் அவசியம் ஆகும்.

மொழியின் தொடரியல் அமைப்புகள் (Syntactic Structure) மூன்று வகைகளாய் பிரிக்கப்படுகின்றன.

1. SVO அமைப்பு

ஜெர்மன், பிரஞ்சு, ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் SOV அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரீஸ், அரேபி, ஷிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருப்புகள் (Preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத் திட்டங்களுக்கு மொழிகளுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரில் *வீடு* என்பது தலைச்சொல் (Head Word) ஆகும். *இராமன்* என்பது சார்புச் சொல் (Dependent Word) ஆகும். தலைச்சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை *-இன்* என்ற ஒட்டு விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையிலான முறையைப் பொறுத்து மொழி தலைக்குறித்தல் மொழி, சார்புக்குறிதல் மொழி எனப் பிரிக்கப்படும். சார்பு குறித்தல் மொழியில் ஒட்டு சார்புச் சொல்லுடனும் தலைக்குறித்தல் மொழியில் ஒட்டு தலைச் செல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி தலைக்குறித்தல் மொழியைச் சார்ந்ததாகும். ஹங்கேரிய மொழி சார்புக்குறித்தல் மொழியைச் சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு விளக்கும்.

English	:	The Man’s house
Hungarion	:	The man house his
House	:	Head word (தலைச்செல்)
The Man	:	Dependent word (சார்புச்சொல்)
-s	:	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியை மேற்கொள்கின்ற போது தலைச்சொல்லுக்கும், சார்பு சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப் பெறுகின்றன என்பது பற்றியெல்லாம் ஆராயவேண்டும்.

மூலமொழியில் உள்ள ஒரு சொல்லின் வகைப்பாடும் இலக்கு மொழியிலுள்ள அதற்கு நிகரான சொல்லின் வகைப்பாடும் ஒன்றாக

இருக்கவேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழி பெயர்ப்பு செய்யப்படும் போது மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்கவேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்புப் பணியில் உள்ள சிக்கல்களாகும்.

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கானச் சொல் விளக்கமானதாகவோ கருக்கமானதாகவோ இருக்கலாம். எடுத்துக்கட்டாக *சகோதரன்* என்ற உறவைக் குறித்து ஆங்கிலத்தில் *brother* என்ற சொல் உள்ளது. மூத்த சகோதரனை ஆங்கிலத்தில் *elder brother* என்று சொல்லுகிறோம். இவ்வாறு வயது மூத்தவர் இளையவர் என்பதைக் குறிக்க *elder*, *younger* என்ற அடைச் சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் *அண்ணன்* என்றும் *தம்பி* என்றும் தனித்தனிச்சொற்கள் உள்ளன. *சாகோதரன்* என்ற உறவை மேலும் வகைப்படுத்த முதியவர், இளையவர் என்ற நிலையில் *அண்ணன்*, *தம்பி* என்ற இருச்சொற்கள் உள்ளன. ஆனால் ஆங்கிலத்தில் இவ்வகையிலான வேறுபாடு என்பது இல்லாதால் உறவை வேறுப்படுத்த *elder*, *younger* என்ற சொற்களைப் பயன்படுத்த நேர்கிறது.

மூலமொழியில் உள்ள செய்திகளில் கலாசாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்கு மொழியின் கலாச்சாரப் பண்புகளுக்கு ஏற்ப மாற்றித்தரத்தக்க அளவு இருத்தல்வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சார தொடர்புகள் மொழிபெயர்ப்பு செய்ய இயலுமா? இயலாதா? என்பதைக் தெளிவுபடுத்த உதவும். மேலும் சபீர்-ஊர்ஃப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ குறைப்பதாகவோ உள்ளது. எனவே மொழிபெயர்ப்பில் இத்தகைய உளவியல் அடிப்படையிலானச் சிக்கல்கள் உள்ளன.

தொடரியல் பகுப்பாய்வி பல பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிபெயர்ப்பிற்காகப் பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்குத் தொடரின் அல்லது வாக்கியத்தின் பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வு செய்யப்படுகிறது. எனவே இயந்திர மொழிபெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது பொருளைக் சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அறிவு அடிப்படை (Knowledge Base) தேவைப்படுகின்றது.

இயந்திரம் வழி மொழிபெயர்ப்பு செய்கின்ற பொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்குக் கூடுதல் கதவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கலைப் புரிந்துக்கொண்டு அதற்கேற்றத் தீர்வு உரிய முறையில் வழங்கப்பட்டவேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்கவேண்டியது அவசியம் ஆகும். பெரும்பாலும் படைப்பாற்றல் திறனைச் செய்திகளில் வெளிப்ப்படுத்தும் பொழுது கணிப்பொறிக்குக் கூடுதல் செய்திகள் தேவைப்படுகிறன்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிபெயர்ப்புகளிலும் உள்ளீடு செய்யும் செய்திகளை முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும் இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிபெயர்ப்பது, பின்னர் மொழி பெயர்ப்பைப் படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகின்றதா என்பதை ஆராய்வது, அப்படி இல்லையெனில் மூலமொழி செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிபெயர்ப்பு செய்வது, சரியான கருத்து வெளிப்படுவதுவரை இத்தகைய ஆய்வை மீண்டும் மேற்கொண்டு மொழி பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து

தேவையான திருத்தங்களைச் செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தை மிச்சப்படுத்துகிறது; மொழிபெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிபெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப்பணித் திறமையானதாக அமைகிறது.

மொழிபெயர்ப்புப் பணிகளுக்காக உருவாக்கப்படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லாது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிறவகை அகராதிகளை விட துறை சார்ந்த அகராதிகள் மொழிபெயர்ப்புப் பணியை எளிமைப்படுத்துகிறது. மூலமொழியில் உள்ள செய்திகளைப் பொறுத்து அது தொடர்புடைய அகராதியைப் பயன்படுத்திக்கொள்ளலாம்.

மொழிபெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத் தொடர்கள் மேலும் அடிக்கடி பயன்படுத்தக்கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்படவேண்டும்.

கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணி என்பது ஒரே மூச்சில் ஒட்டு மொத்தமாகச் செய்யக்கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்படவேண்டும். அதன் செயல் திறனை அறியவேண்டும். பணியின் தன்மையை மதிப்பிடவேண்டும். இதற்கு ஏற்றார் போல் தேவைப்படின் மென்பொருளைக் கூடுதல் வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை மனிதனுக்கு நிறைவழிக்கும் கொள்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்புக் கடுமையானப் பணியாகும். இதற்கு மனிதமொழி பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (Model)

வழிமுறைகளும் அவசியமாகும். கணிப்பொறி வழி மொழிபெயர்ப்பு எண்ணற்ற முயற்சிகளால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமான பணியாகும். இதற்கு அறிவியல் திறன் மட்டுமல்லாது மொழிபெயர்ப்பிற்குத் தேவையான கலைத்திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் நடைபெற்றுள்ள மொழிபெயர்ப்புப் பணிகள் மனதுக்கு நிறைவழிக்கக்கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்கு உதவுகின்றன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்குப் போதுமானதாகும்.

மனிதனின் படைப்பாற்றல் கற்பனைத்திறன், உளவியல்பாங்கு போன்ற செய்திகளால் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியில் இத்தகையக் கூறுகளைச் சேர்க்கவேண்டி உள்ளது. அதற்கு மனித நுண்ணிறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன.

2.7 இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி

ஆரம்பக்கட்டத்தில் இந்தியாவில் இயந்திர மொழிபெயர்ப்புகள் இதற்கான சாத்தியத்தை வெளிப்ப்படுத்தும் பொம்மை ஒழுங்குமுறைகளாக (Toy Systems) இருந்தன. 1978 வரை இயந்திர மொழிபெயர்ப்பிற்கான எந்த முயற்சியும் எடுக்கப்படவில்லை. 1978-இல் ஐ.ஐ.டி.கான்பூரில் உள்ள மின்னியல் தொழில் நுட்பத் துறையின் முயற்சியால் மொழியியல் அடிப்படையாலன தகவல் ஒழுங்குமுறைகள் பற்றிய தேசியக் கருத்தரங்கு (National Symbosium on Linguistic Based Information System) முதன் முதலில் நடைபெற்றது. இக்கருத்தரங்கு ஆராய்ச்சியாளர்களிடையேயும் அரசாங்கத்திலும் தொழில்நுட்ப துறைகளிலும் இயந்திர மொழிபெயர்ப்பு குறித்த விளிப்புணர்ச்சியை ஏற்படுத்தியது. இருப்பினும் இயந்திர மொழிபெயர்ப்பு 80-களின் ஆரம்பகட்டத்தில் இந்திய மொழிகளுக்குச்

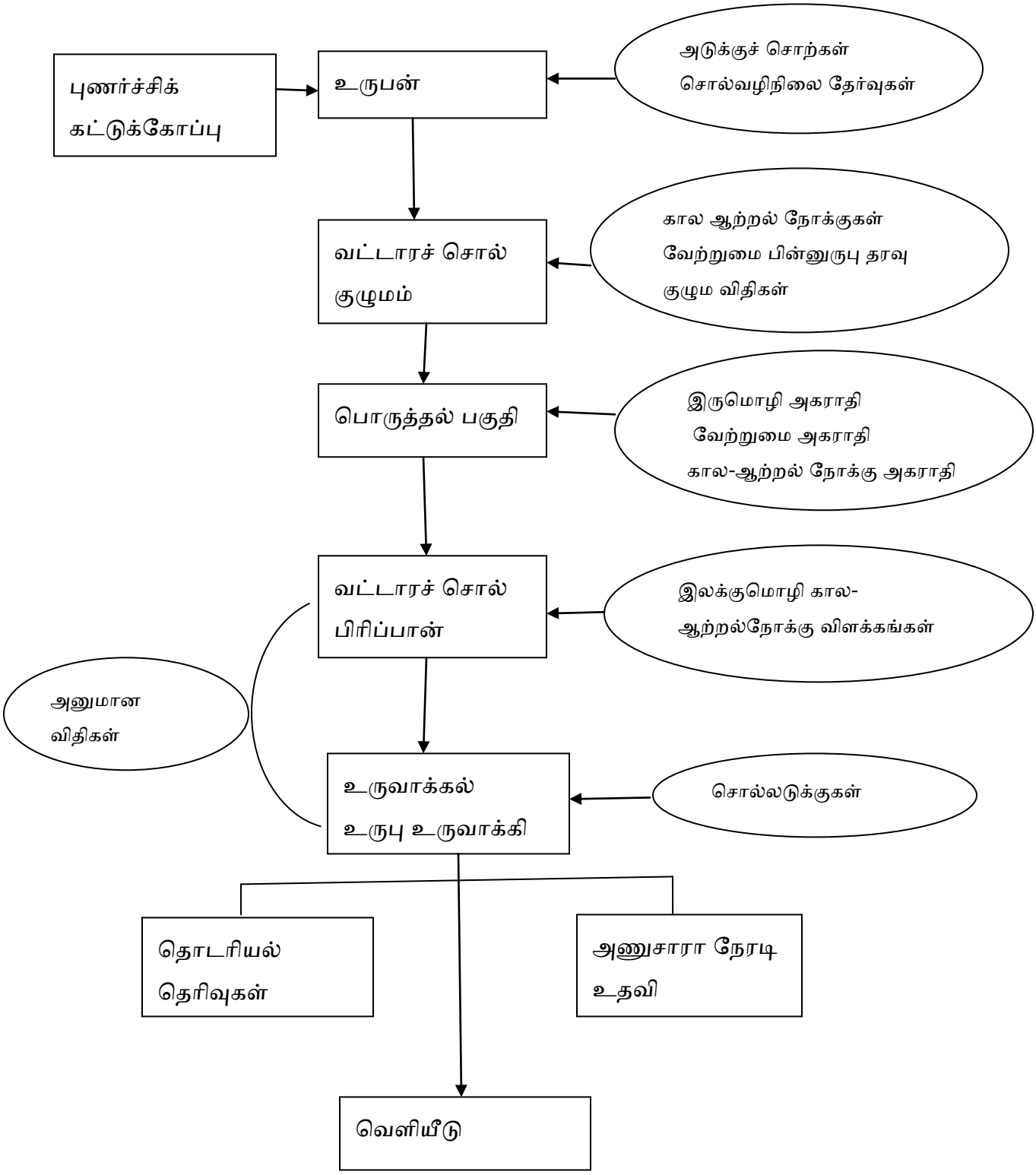
செல்லாய்விகளும் பிற கணிபொறி வசதிகளும் வரத் தொடங்கிய பின்னர்தான் தொடங்கப்பட்டது. ஐ.ஐ.டி.கான்பூர் GIST என்ற தொழில் நுட்பத்தின் உருவாக்கத்திற்கு முயன்றது. இது இயந்திர மொழிபெயர்பிற்கு விழிப்புணர்ச்சி ஏற்படுத்துவதுடன். இதில் ஆய்வு மற்றும் வளர்ச்சி குறித்த பல திட்டங்களை ஊக்குவித்தது. இந்தியாவில் பல மையங்களில் பலவிதமான முயற்சிகள் மேற்கொள்ளப்பட்டாலும் மூன்றும் அணுகுமுறைகளைப் பட்டியலிட இயலும்.

1. இடமெழி அணுகுமுறை (Interlingual Approach)
2. நேரடி சொல்சார் மாற்றம் (Direct lexical Transfer Approach)
3. கணிப்பொறி உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிகள் (Machine Aided Translation Effort)

2.7.1 இடைமொழி அணுகுமுறை (Interlingual approach)

சமஸ்கிருதத்தை இடைப்பட்ட அடிப்படை மொழியாகக் கொண்டு இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பு செய்யும் இடைமொழி ஆய்வு அணுகுமுறை சின்ஹா என்பவரால் 1984-இல் முயற்சிக்கப்பட்டு 1989-இல் விரிவாக்கப்பட்டது. இது மூல மொழியின் ஆய்வுக்குக் காரக அடிப்படையிலான அக உருப்படுத்தத்தை முன்மொழிந்தது. 1986-1988-களில் எல்லைகுட்பட்ட சொற்களைக் கொண்ட மிக எளிய வாக்கியங்களை இந்தியிலிருந்து தெலுங்கிற்கு மொழிபெயர்க்கும் மூலமுன்மாதிரி (prototype) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் உருவாக்கப்பட்டது. இது காரக அடிப்படையிலான கருத்துவரைபடத்தை அக உருப்படுத்த்திற்கும் பொருண்மை மயக்கங்களுக்கும் பயன்படுத்தியது. ஒரு தனிநிலை தொடரியல் இருப்பானில் உள்ள மிகக்கூடுதலாகத் தொடர்புள்ள சொற்கள் ஒரு சொல் குழுவாகப் பயன்படுத்தப்பட்டது. மூல மொழியில் உள்ள உரை உருவாக்கி இச்சொற்

குழுமங்களிலிருந்து பொருத்தத்தைச் சரியான உருபனியல் உருவாக்கத்திற்குப் பயன்படுத்தியது.



இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயரிப்பு அதற்குப்பின் முயற்சிக்கப்படவில்லை. இவ்வணுகுமுறை இந்திய மொழிகளின் பன்மைத்தன்மைக்குக் கூடுதல் அனுகூலமாய் இருந்தாலும் அக உருப்படுத்ததிலிருந்து புற அமைப்பை பெறுவது கடினமான செயல்பாடாக அமைந்து குறையுடையதாய் இருந்தது. இவ்வணுகுமுறை மூலமொழியின் சீரான புரிதலை முன்கருதலாகக் கொண்டதால் இன்றைய ஆய்வுநிலையில் ஏற்றதாக அமையவில்லை. மேலும் அமைப்பு அடிப்படையில் அண்மைப்படுகின்ற இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பின் போது நேரிடையான சொல் பதிலீடு செய்வது எளிமையாக அமைந்தது. இதன் காரணமாக இவர்களால் அனுசாரகா என்ற அணுகுமுறை உருவாக்கப்பட்டது.

இந்திய மொழிகளுக்கிடையே உள்ள அமைப்பு ஒற்றுமை காரணமாக மூலமொழியிலிருந்து இலக்கு மொழிக்குப் பல பொருண்மை மயக்கங்களைக் கொண்டு செல்ல இயலும் என்ற காரணத்தால் இந்த இடைமொழி அணுகுமுறை இந்திய மொழிகளுக்கிடையே மொழிபெயர்ப்பிற்கு நல்லதொரு அணுகுமுறையாக அமையும். இருப்பினும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து ஆங்கில மொழிகளுக்கும் மொழிபெயர்க்க இடைமொழி அணுகுமுறை பொருத்தமான அணுகுமுறை அல்ல.

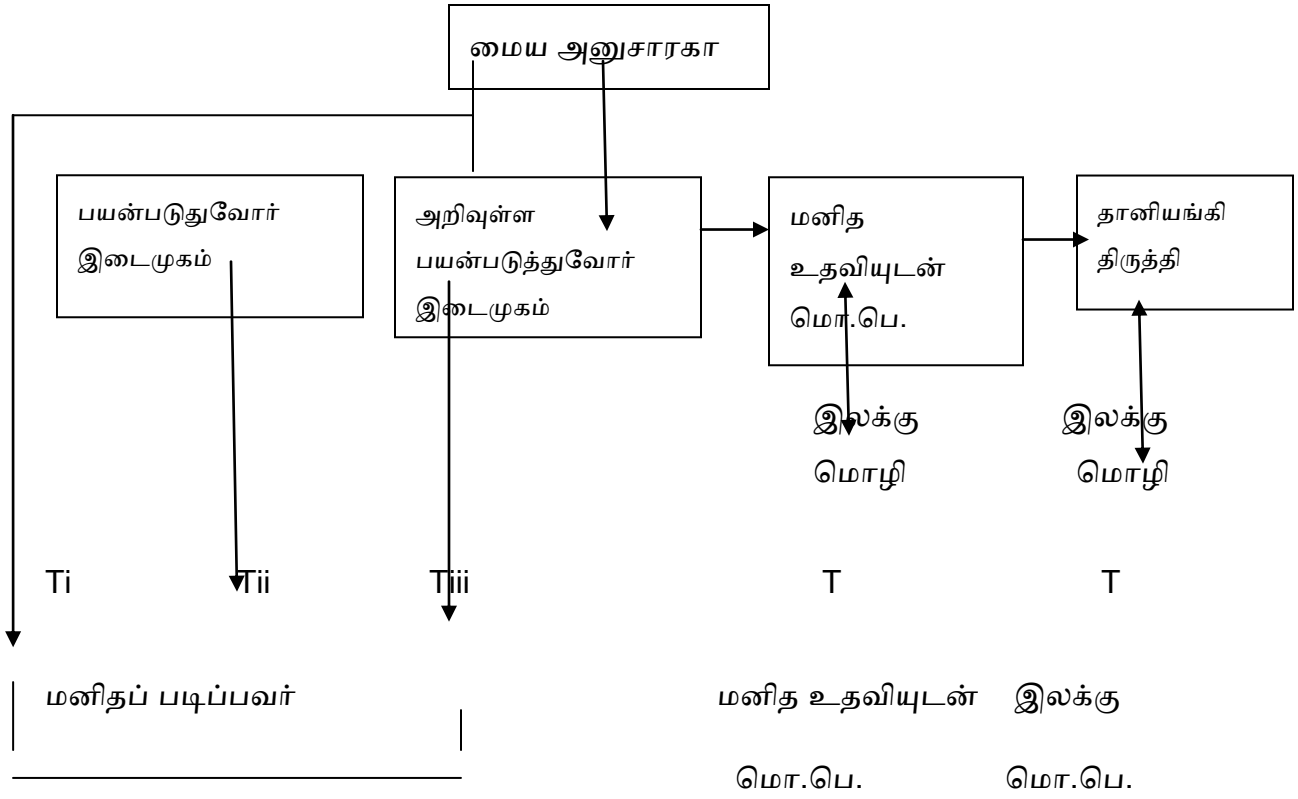
2.7.2 சொல்நிலையில் மாற்றத்தைப் பயன்படுத்திய நேரடி அணுகுமுறை (அனுசாரக் திட்டம்)

சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் மூலமொழியிலிருந்து இலக்கு மொழிக்கு நேரடி சொல்சார் மாற்ற அணுகுமுறை உருவாக்கப்பட்டது. இங்கு மூல மொழிக்கும் இலக்கு மொழிக்கும் இடையில் உள்ள வாக்கிய அமைப்பின் பொதுமைப் பண்புகள், இரண்டும் இந்திய மொழிகள் என்ற காரணத்தால் கருதப்பட்டு முழு அளவிலும் பயன்படுத்த்ப்பட்டுள்ளது.

இவ்வணுகுமுறையின் படி ஒரு தனிநிலை தொடரியல் உறுப்பாக ந்டத்தை செய்யும் மூலமொழியிலுள்ள சொற்கள் குழுமப்பட்டு மூலமொழிக்குப் பொருந்தும்படி அது சொற்களால் பதிலீடு செய்யப்படும் பீன்னுருபை நிர்ணயிக்க சில் விதிகள் உருவாக்கப்பட்டன. சொற்களின் நிரல் வருகை பெரும்பாலும் தக்க வைக்கப்பட்டு சில சமயங்களில் புதிய இணைப்பான்களின் பயன்பாட்டால் நிறைவு செய்யப்படும். சில சமயங்களில் இம்முறை வெளியீட்டில் தொடரியல் தவறுகளுக்கு வழிவக்கும்; இது பற்றி இதை உருவாக்கியவர்கள் கவலைப்படவில்லை. இருப்பினும் குறைந்த அளவு பொருள் பெரும்பாலும் வெளிப்படுத்தப்படும். மேலும் அனுசாரக்காவின் பாகங்கள் அல்லது உறுப்புத் தொகுதிகள் எந்த இயந்திர மொழிபெயர்ப்பிற்கும் அல்லது இயந்திர உதவியுடன் செய்யப்படும் மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கும் பகுதியாக அமையும்.

இவ்வாறு அனுசாரகாவின் வெளியீடு இடைப்பட்ட ஒரு நிலைக்கு முக்கிய வெளியீடாகும் மற்றும் அனுசாரகா போன்ற வெளியீடு ஒரு இரண்டாம் நிலை விளைவாகும் (by product). இது சீரற்ற மொழிபெயர்ப்பிற்கு கொண்டுசெல்லாததால் மொழிகளின் இணையின் இயந்திர மொழிபெயர்ப்பிற்கும் பயனுள்ள உபாயமாகும்.

இதை உருவாக்கியவர்கள் இதை ஒரு மொழி அணுகி(language acessor) என்று அழைகின்றனர்; இதிலிருந்து முழு நன்மை அடைய பயன்படுத்துபவர்களிடம் சில் பயிற்சியை எதிர்பார்க்கின்றனர். தொடக்கத்தில் கன்னடா, இந்தி மொழிகளுக்கு ஒரு அனுசாரக் ஒழுங்குமுறை அமைக்கப்பட்டு காட்டப்படது. இதன் பின்னர் தெலுங்கு-இந்தி, பஞ்சாபி-இந்தி, மராட்டி-இந்தி, சமஸ்கிருதம்-இந்தி, தமிழ்-இந்தி ஆகிய மொழி இணைகளுக்கு அனுசாரக் ஒழுங்குமுறைகள் உருவாக்கப்பட்டு பார்வையாளர்களுக்கு எடுத்துக்காட்டப்பட்டது.



அனுசாரகா வெளியீட்டில் \

பல மட்டங்கள்

2.7.3 ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு இயந்திர உதவியுடன் மொழிபெயர்ப்பு (ஆங்கிலபாரதி)

ஐ.ஐ.டி.கான்பூரில் மேற்கொள்ளப்பட்ட ஆங்கில பாரதி (*ANGLABHARATI*) திட்டத்தை நிறுவுதில் நான்கு முக்கிய கருதல்கள் இருந்தன. முதலாவது அன்றைய காலகட்டத்தின் நிலையில் சீரான இயந்திர மொழி பெயர்ப்பு சாத்தியமில்லை. எனவே இயந்திர உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிக்கப்பட வேண்டும்; இங்கு இயந்திரம் பெரும்பாலான செயல்களைச் செய்கின்றது. கடினமான வேலைககள்

முந்தைய மற்றும் பிந்தைய திருத்தங்களால் (Pre and Post editings) கையாளப்படுகின்றது. இதன்படி ஒவ்வொரு மூலமொழி வாக்கியத்தையும் தனியாகப் பரிசோதிப்பது மற்றும் முற்கட்டு மற்றும் (anaphora) வாக்கியங்களுக்கிடையிலான குறிப்புகள் (Intersentential References) போன்ற சிக்கல்களை மனிதத் திருத்துனருக்கு விட்டுவிடுவது என்பன எளிமைக்கு வழி வகுக்கிறது. இரண்டாவது இயந்திர மொழி பெயர்ப்பு, பயனப்படுத்துவோர் மற்றும் பயன்பாட்டு அடிப்படையில் இருக்கவேண்டும் என்ற உணர்வு, ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு உள்ள மொழிபெயர்ப்பை மேற்கொள்ளுவது இந்திய மொழிகளுக்கு இடையிலான இயந்திர மொழிபெயர்ப்பை மேற்கொள்வதைக் காட்டிலும் சிறந்தது என்ற தீர்மானத்திற்கு வழிவகுத்தது. மூன்றாவது ஒரு குறிப்பிட்ட பொருண்மைக்களத்தை அடிப்படையாகக் கொண்ட ஒழுங்குமுறைகள் (domain specitic systems) திறந்த ஒழுங்குமுறைகளைக் காட்டிலும் (open ended systems) உருவாக்குவதற்கு எளியது மற்றும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு பயன்படுத்துவோர்க்கு சிக்கனமானதாகவும் இருக்கிறது. இறுதியாக வேறுப்பட்ட மொழிகள் ஒற்றுமையுள்ள மொழிகளைக் காட்டிலும் ஆய்வு அடிப்படையில் கூடுதல் அறை கூவல்களை எதிர்கொள்ளும்.

சில முக்கியத்திட்டக் கருத்தல்கள் 90 விழுக்காடு வேலையை இயந்திரத்தாலும் வேலையை மனித முன் திருத்தத்தாலும் (post editing) மேற்கொள்ளும் படிக்கு மொழிப் பெயர்ப்பிற்கு ஒரு பயன்பாட்டு கருவியைத் தருவது என்ற நோக்கம் அடிப்படையில் அமைந்தது, இந்த அமைப்பொழுங்கு படிப்படியாகச் சிக்கலான சூழல்களைக் கையாளும்படிக்கு வளர இயலும், பொத்துமான உரை உருவாக்கும் பகுதிகளின் (text generator modules) இணைப்பால் ஆங்கிலத்திலிருந்து பெரும்பாலான இந்திய மொழிகளுக்கு மொழிப்பெயர்ப்பு ஒரு சீரான இயங்கு முறையைக் கொண்டிருந்தது. மேலும் அதன்

பயன்பாட்டிற்கும் விரிவாகத்திற்கும் வசதி செய்ய மனிதனால் இயக்கப்பட்ட மனிதன் - இயந்திர இடைமுகம் கொண்டிருந்தது.

இது ஒரு குழும இந்திய மொழிகளுக்கு (target languages) பயன்படுமாறு போலி இலக்குமொழியை உருவாக்கும் (Pseudo target language) சூழல் வரையறையற்ற இலக்கணம் போன்ற அமைப்புடன் கூடிய அமைப்பொழுங்கை இலக்காகக் கொண்ட விதி அடிப்படையிலான ஒழுங்குமுறையாகும். பெருந்தரவு ஆய்வின் மூலம் கிடைக்கப்பெற்ற ஒரு குழும விதிகள் போலி மூலமொழிக்கு நகர்வு விதிகளைப் பயன்படுத்தும் சாத்தியமான உறுப்புகளைக் கண்டு கொள்வதற்குப் பயன்படுத்தப்படுகிறது. போலி இலக்குமொழியைப் பயன்படுத்தும் கருத்து, இடைமொழி அணுகுமுறையைப் பயன்படுத்துவதற்குச் சமமான நன்மையைப் பெற அமைப்புசார் ஒற்றுமையை முதன்மையாகப் பயன்படுத்துவதாகும். போலி இலக்குமொழி, இடைமொழி அணுகுமுறையில் பயன்படுத்தப்படும் இடைப்பட்ட மொழி அல்ல இங்கு பொருண்மை உருப்படுத்தத்தை உருவாக்க எந்த முயற்சியும் மேற்கொள்ளப்படவில்லை.

இந்திய மொழிகள் சுதந்திர சொல்குழும நிரலைக் கொண்டு வினையில் முடியும் பண்புடையனவாகும். அமைப்பு ஒற்றுமையின் வழி இந்திய மொழிகளை அவற்றின் மூலம் அடிப்படையில் நான்கு பெரிய குழுமங்களாக வகைப்படுத்தலாம்:

1. இந்திய – ஆரிய மொழிக்குடும்பம் (Indo-Aryan family) (இந்தி, வங்காளம் அஸ்ஸாமிஸ், பஞ்சாபி, மராத்தி, ஒரியா குஜாத்தி போன்றன)
2. திராவிட மொழிக் குடும்பம் (Dravidian Family) (தமிழ், தெலுங்கு, கன்னடம், மலையாளம் போன்ற)
3. ஆஸ்ட்ரோ – ஆசிய மொழிக் குடும்பம் (Astro-Asian family),
4. திபெத்திய பர்மிய மொழிக் குடும்பம் (Tibato-Burman Family).

ஒவ்வொரு குழுமங்களுக்குள்ளும் மொழிகள் உயர்ந்த அளவு அமைப்புப் பொருத்தத்தைக் காட்டுக்கின்றன. ஆங்கில பாரதி திட்டம் இந்த ஒற்றுமையைக் கூடுதல் அளவு இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கப்பயன்படுத்துகிறது.

இலக்குமொழியில் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கு வேண்டி பல பொருண்மை அடையாளங்கள் பயன்படுத்தப்படுகின்றன. தீர்க்கப்படாத பொருண்மை மயக்கங்களின் மாற்றுப் பொருள்கள் போலி இலக்குமொழியில் தக்க வைக்கப்படுகிறது. ஒவ்வொரு இலக்குமொழிக்கும் உரை உருவாக்கப் பகுதி போலி இலக்குமொழியை இலக்குமொழிக்கு மாற்றுகிறது. ஓவ்வொரு இலக்குமொழிக்கும் திருத்துவான் (corrector) பயன்படுத்தப்படுகிறது. இறுதியாக மனித இயக்கத்தால் தூண்டப்பட்டபின் திருத்தும் தொகுதி (Post editing package) இறுதி திருத்தங்களைச் செய்வதற்குப் பயன்படுத்தப்படுகிறது. பின்திருத்தி இலக்குமொழியை மட்டும் அறிந்தால் போதுமானது.

ஆங்கில பாரதி திட்டம் சின்ஹா (Sinha, 1993) என்பவரால் 1991-ல் உருவாக்கப்பட்டது. இத்திட்டத்தில் ஆங்கிலத்திலிருந்து இந்தி மற்றும் தெலுங்கு மொழிகளுக்குச் செயல்படக்கூடிய மூலமுன்மாதிரி மொழிபெயர்ப்பு (Functional Prototype) ஒழுங்குமுறை உருவாக்கப்பட்டது. பின்வரும் திட்டவரைபடம் இந்த மொழி பெயர்ப்பு ஒழுங்குமுறையை விளக்கும். பின்வருவன ஆங்கில பாரதி மொழிபெயர்ப்புத் திட்டத்தின் முக்கியக்கூறுகளாகும்.

1. விதி அடிப்படை

விதி அடிப்படை (Rule base) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு வாக்கியங்களின் அமைப்புகளைப் பொருத்துவதற்கான விதிகளைக் கொண்டிருக்கின்றது. இந்த ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு அமைப்பு

ஒழுங்கு மாற்றங்களின் தரவு மையம் மொழிபெயர்களுக்கு அமைப்பு ஒழுங்கு மாற்றங்கள் தரவு மையம் மொழிபெயர்க்கவேண்டிய வாக்கியத்திற்குகாகக் கிளை அமைப்பைப் பெறும் வேலையைப் புறக்கணித்து புறக்கிளைகளிலிருந்து புறக்கிளை மாற்றத்தை செய்யும் வேலைக்குப் பணிக்கப்படுகிறது. வாக்கியங்கள் பொருண்மையைக் கண்டு பிடிக்க புற அமைப்பொழுங்கைப் பயன்படுத்தும் கருத்து, மொழியியலில் மிகப் பழமையானதாகும். இந்த அணுகுமுறை எளிதானது என்றாலும் இத்திட்டம் ஒரு மொழியின் புற அமைப்பு ஒழுங்களின் தனிப்பட்ட தன்மைகளைக் கண்டுபிடிக்கச் செயலூக்கமுடையதாகும். ஆங்கில பாரதியில் பயன்படுத்தப்படும் அணுகுமுறை சாம்ஸகியின் தொடரமைப்பு இலக்கணத்திலிருந்தும் சொல்சார்- செயல்பாடு சார் இலக்கணத்தின் (Lexical functional Grammer) அமைப்பிலிருந்தும் (C-structure) அதிகமாக எடுத்தாண்டுள்ளது. விதி அடிப்படை (rule base) என்று கூறப்படும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கான அமைப்பு மாற்ற விதிகளின் தரவு மூலம் ஆங்கில பாரதி ஒழுங்குமுறையில் மையப் பகுதியாக அமைகிறது. இது ஆங்கிலத்திலிருந்து மொழிபெயர்கும் போது தொடரியலில் ஏற்படும் முக்கிய மாற்றங்களைக் கவனித்துக்கொள்கிறது. முன்னர் கூறியப்படி இந்திய மொழிகளுக்குப் பொது விதி அடிப்படையை உருவாக்கி ஆங்கில பாரதி ஆங்கிலத்திலிருந்து மொழிபெயர்க்கும் போது சிறந்த நன்மையை வெளிப்படுத்துகிறது.

2. அர்த்த மயக்க நீக்கி

அர்த்த மயக்கநீக்கி (Sense disambiguator) மூல மொழியின் ஒவ்வொரு சொல்லுக்கும் சரியான அர்த்ததை தேர்ந்தெடுக்கப் பொறுப்புள்ளதாகும். இங்கு அர்த்த மயக்கங்க நிக்கம் மூல மொழியின் உரையில் மட்டுமே நிகழ்த்தப் பெறுகின்றது. ஆங்கில பாரதியில் பயன்படுத்தப்படும் இவ்வணுகுமுறையை விதியால் விதி-பொருகோள் (Rule-by-rule Semantic Interpretation) எனக்

கூறலாம். ஒரு தொடரியல் விதி பயன்படுத்தப்படும் ஒவ்வொரு நேரத்திலும் பொருள்கோளி (Semantic Intepreter) பயன்படுத்தப்படுகின்றது.

3. இலக்கு உரை உருவாக்கி

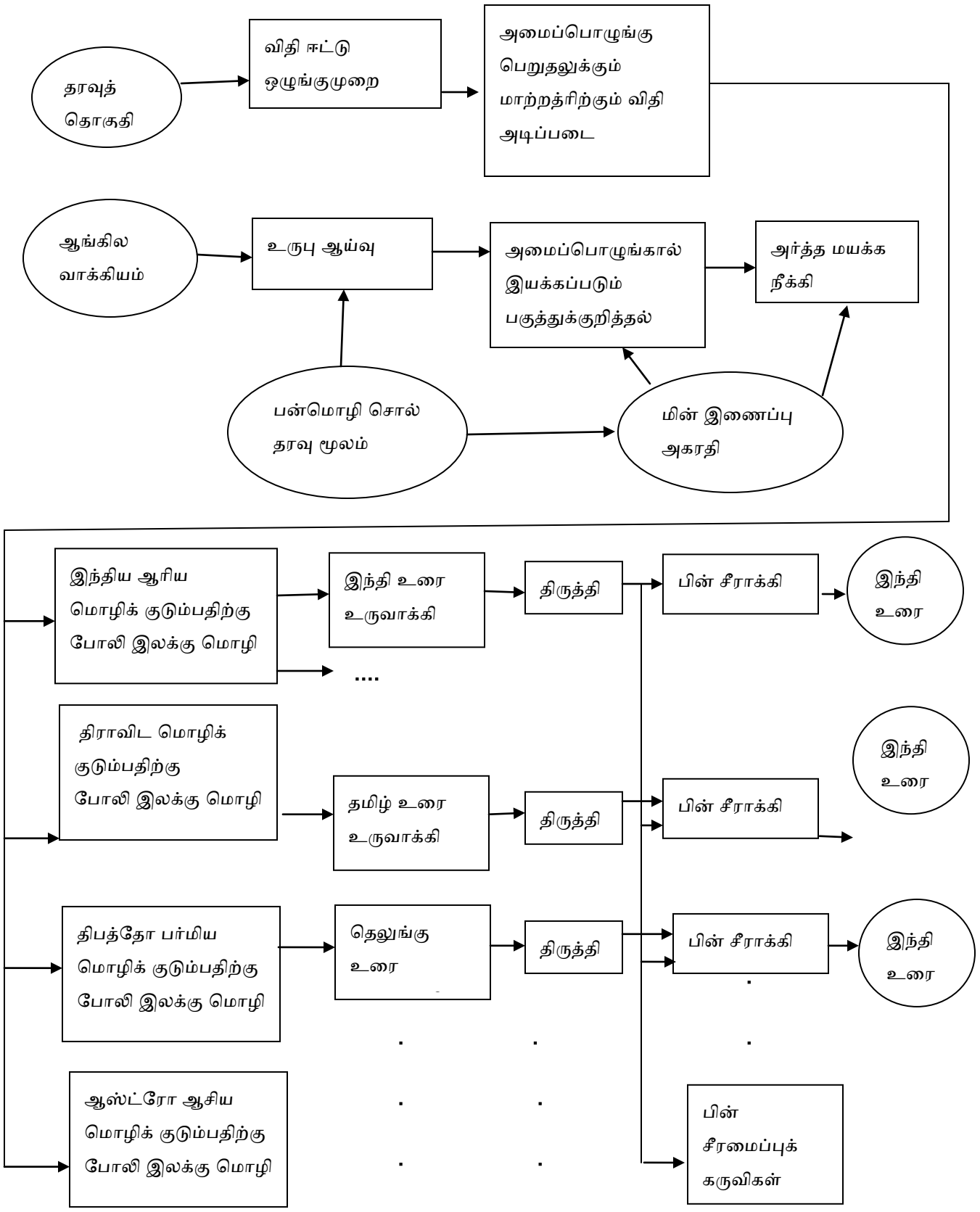
இலக்கு உரை உருவாக்கி (Target text generator) மொழிபெயர்ப்பு ஒழுங்குமுறையின் இறுதியாக அமைகின்றது. இதன் செயல்பாடு பொறுத்தமுறும் மூல மொழிகளிலிருந்து மொழிபெயர்க்கப்பட்ட வெளியீட்டை உருவாக்குவது ஆகும். இது ஆங்கில பாரதியின் முந்தைய நிலைகளின் உருவாக்கப்பட்ட இடைப்பட்ட வடிவத்தை உள்ளீடாக ஏற்றுக் கொள்கின்றது. இவ்வேலையை இயற்கை மொழி உருவாக்கம் (Natural Language Generation) என்று அழைக்கப்படும் வேலையிலிருந்து வேறுபட்டவையாகும்; இவ்வர்த்ததில் பிந்தையது எனக் கூறவேண்டும் என்பதுடன் திட்ட நிலை (Strategic level) எவ்வாறு கூற வேண்டும் என்பதையும் நடவடிக்கை நிலை (Tactical Level) தீர்மானிக்க வேண்டும். ஒரு விதி அடிப்படை மற்றும் அர்த்த மயக்க நீக்கி இவற்றைப் பயன்படுத்தும் வேறுபட்ட உரை உருவாக்கிகளைக் கொண்டு பல இலக்கு மொழிகளுக்கும் பொது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை அமைப்பு கிடைக்கப்பெறுகின்றது.

4. பன்மொழி அகராதி

பன்மொழி அகராதி (Multilingual Dictionary) தொடரியல் வகைப்பாடுகள், சாத்தியமான அர்த்தங்கள், அர்த்தங்களின் மயக்கத்தை நீக்கும் முக்கிய கூறுகள் (Keys) இவற்றை உள்ளடக்கிய ஆங்கிலச் சொற்களுக்கும் இணையாக மூல மொழிச் சொற்களுக்குப் பல விளக்கங்களைக் கொண்டிருக்கும்.

5. விதி அடிப்படை ஈட்டி

விதி அடிப்படை ஈட்டி (rule base acquirer) மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு விதி அடிப்படையை உருவாக்குகிறது.


தரவுத் தொகுதி
விதி ஈட்டு ஒழுங்குமுறை
அமைப்பொழுங்கு பெறுதலுக்கும் மாற்றத்திற்கும் விதி அடிப்படை
ஆங்கில வாக்கியம்
உருபு ஆய்வு
அமைப்பொழுங்கால் இயக்கப்படும் பகுத்துக்குறித்தல்
அர்த்த மயக்க நீக்கி
பன்மொழி சொல் தரவு மூலம்
மின் இணைப்பு அகரதி
இந்திய ஆரிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
இந்தி உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
....
திராவிட மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தமிழ் உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
திபத்தோ பர்மிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தெலுங்கு உரை
திருத்தி
பின் சீராக்கி
இந்தி உரை
ஆஸ்ட்ரோ ஆசிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
பின் சீரமைப்புக் கருவிகள்

2.8 தமிழ் சார் இயந்திர மொழிப்பெயர்ப்பு

தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள் பல மேற்கொள்ளப்பட்டு தமிழி நாட்டில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சிக்கு வித்திட்டன.

2.8.1. தமிழ் சார் இயந்திர மொழிபெயப்பு முயற்சிகள்

2.8.1.1 தமிழ் உருஷ்யன் மொழிப் பெயர்ப்புத்திட்டம்

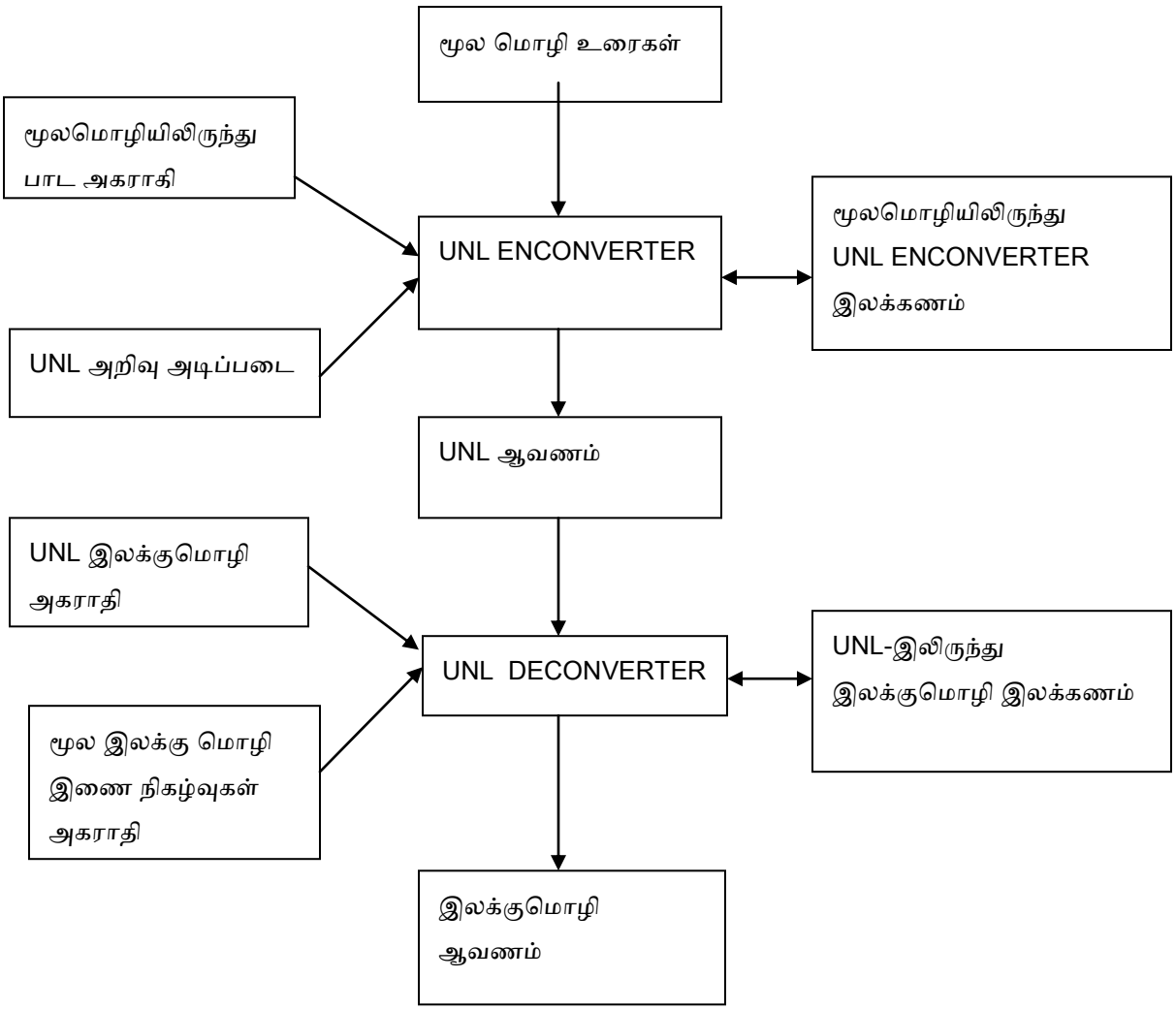
1985-ஆம் தமிழ் பல்கலைக்கழகத்தில் (Tumtst Tamil University Machine Inaslation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. இதன் மூலம் ருசஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிப்பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழகம் கண்ணி அறிவியல்துறை, மொழியியல் துறை மற்றும் மொழிப் பொயர்ப்புத் துறை ஆகியவைகளின் கூட்டு முயற்சியில் உருவானதாகும்.

2.8.1.2 உலக வலைப்பின்னல் மொழி – தமிழ்க்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

இந்த அணுகுமுறை சென்னையில் உள்ள அண்ணா பல்கலைக்கழகத்தில் கணிப்பொறி மொழியியல் துறையில் மேற்கொள்ளப்பட்டது. (Dhanabalan & Geetha, 2004). இந்த அணுகு முறையில் இடையீட்டு மொழியின் உருப்படுத்தம் பயன்படுத்தப்படுகிறது. மூலமொழி அகராதியையும் இலக்கணத் தகவல்களையும் பயன்படுத்தி இடைமொழி உருப்படுத்தம் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிப்பெயர்ப்பு செய்ய பயன்படுத்தப்படும் இடைமொழி அமைப்பு உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)) ஆகும். இம்மொழி இடைமொழி அமைப்பாகத் திட்டமிடப்பட்டு மூலமொழியிலிருந்து

இலக்கு மொழிக்கும் இலக்கு மொழியிலிருந்து மூலமொழிக்கும் தானியங்கி இயந்திர மொழிபெயர்ப்புயைச் சாத்தியமாக்குகின்றது. UNL-ஐ பயன்படுத்திய மொழிபெயர்ப்பு ஒழுங்குமுறை மூலமொழியிலிருந்து UNL-க்கு மாற Encoverter-ஐயும் UNL-இலிருந்து இலக்கு மொழிக்கு மாற Deconverter பயன்படுத்துகின்றது.

இந்த அணுகுமுறையின் திட்டவடிவம் கீழே தரப்பட்டுள்ளது.



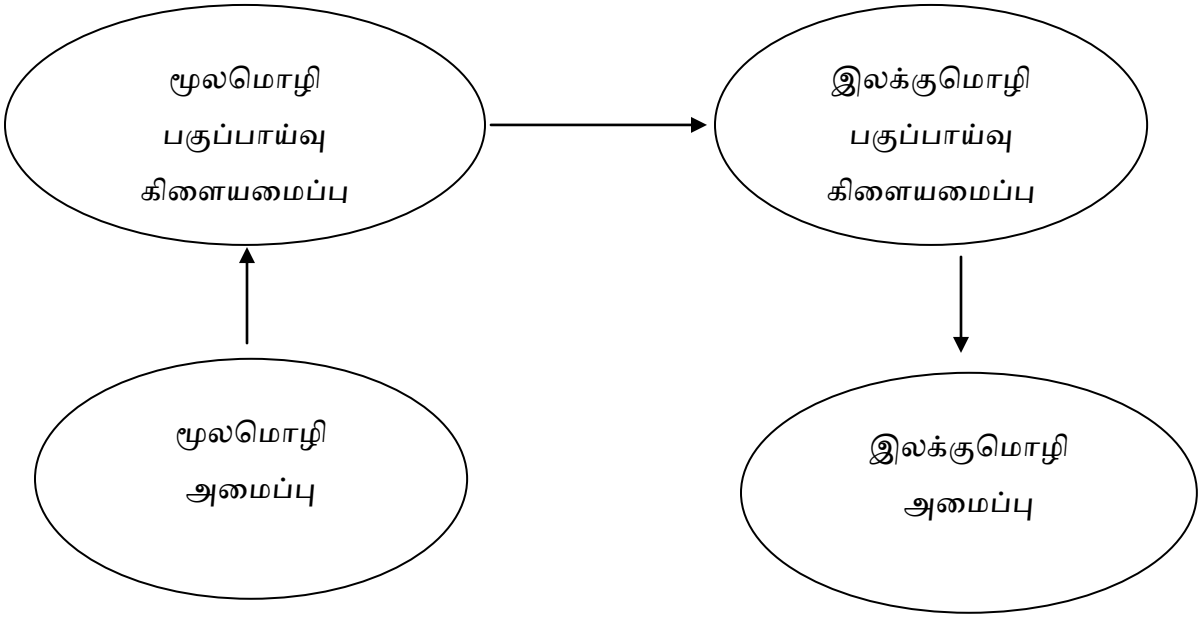
மேற்கண்டவாறு படத்தில் காட்டியவாறு UNL அமைப்பு வேலை செய்து UNL மொழியிலிருந்து தமிழுக்கு மொழிபெயர்ப்பு செய்யும்.

2.8.1.3 ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

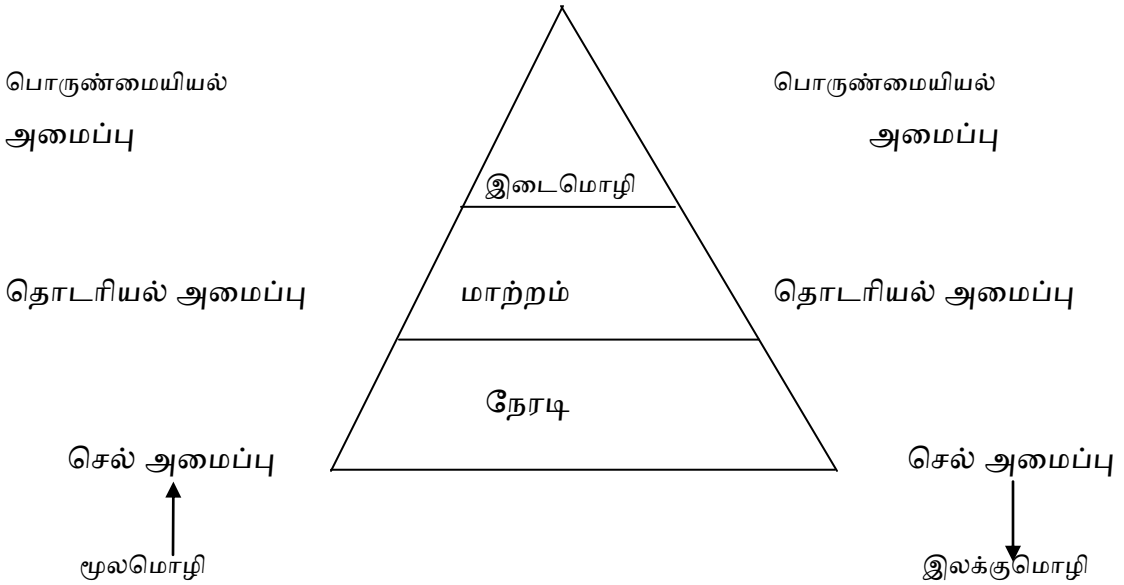
ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்க ஒரு இயந்திர மொழி பெயர்ப்புக் கருவி உருவாக்க, காமாட்சி மற்றும் ராஜேந்திரன் (Kamachi Rajendran, 2004) ஒரு திட்டத்தை உருவாக்கி செயல்படுத்த முயன்றனர். தமிழில் முதுகலை மொழியியல் பாடத்திட்டத்திற்கு வேண்டிய மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்க வேண்டிய கட்டாயம் ஏற்பட்டதால் இத்தகைய ஆய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வு மொழியியல் புத்தகங்களை ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்க உதவும் ஒரு கருவி தயாரிப்பதை நோக்கமாக கொண்டது. இக்கருவி பின்வரும் நான்கு பகுதிகளைக் கொண்டது.

1. தமிழ் மொழி அமைப்புப் பகுப்பாய்வு
2. ஆங்கில மொழி அமைப்புப் பகுப்பாய்வு
3. ஆங்கில - தமிழ் மொழிபெயர்ப்பு அகராதி
4. ஆங்கில – தமிழ்மொழி மாற்றமைப்பு விதிகள்
5. ஆங்கில மொழியமைப்பு உருவாக்கி

இந்த மொழிப்பெயர்ப்பு கருவியின் செயல்பாட்டை விளக்கும்



ஒரு ஒழுங்குமுறையை உருவாக்குவதில் ஏற்படும் சிக்கலின் நிலை மொழி பெயர்க்கப்படும் இரு மொழிகளின் ஒற்றுமை மற்றும் வேற்றுமைகளை பொறுத்து அமையும். இவ்வொற்றுமை வேற்றுமைகள் சொல்நிலையில் மொழி மாற்றம் செய்யப்பட வேண்டுமா? தொடர் நிலையில் மாற்றம் செய்யப்பட வேண்டுமா? பொருண்மை நிலையில் மாற்றம் செய்யப்பட வேண்டுமா என்பதைத் தீர்மானிக்கும் பின்வரும் படம் இச்சாத்தியங்களை வெளிப்படுத்தும்.



தமிழிலிருந்து பிற மொழிகளுக்கு மற்றும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு இயந்திர உதவி மொழிபெயர்ப்பு ஒழுங்குமுறைகள் தமிழ்மொழியின் மீது பல நிலைகளிலான ஆய்வுகளை உள்ளடக்கும். ஒவ்வொரு நிலைக்கும் அதற்கான சிக்கல்கள் காணப்படும்.

2.8.2 தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்

2.8.2.1 உருபனியல் பகுப்பாய்வு (Morphological Analysis)

இது வேர் சொற்களைக் கண்டுபிடிப்பதையும் அதன் இலக்கண பண்புப் கூறுகளைக் கண்டுபிடிப்பதையும் உள்ளடக்கும். தமிழ் ஒரு திரிபு (Inflection) வளமுள்ள மொழி ஆகையால் அதன் பகுப்பாய்வு ஆங்கிலத்தின் உருபனியல் பகுப்பாய்வை விடச் சிக்கல் வாய்ந்ததாக அமையும். புணர்ச்சி விதிகள் தமிழில் உருபனியல் ஆய்வி உருவாக்குவதை கடினமாக்குகின்றது. சொற்களை உருபன்களாகப் பிரிப்பதிலும் அவற்றிற்குப் பொறுத்தமான பொருண்மை மற்றும் இலக்கணத் தகவல்கள் தருவதிலும் கவனம் தேவை. பின்வரும் எடுத்துக்காட்டு இதைக் தெளப்வுப்படுத்தும்.

தமிழ்சொல் : பணத்திற்காகத்தானே

உருபனியல் பிரிப்பு : பணம் + க்கு + ஆக + தான் + ஏ

வேர் சொல் : (பணம்)

நான்காம் வேற்றுமை உருபு (க்கு)

கொடை வேற்றுமை உருபு (ஆக)

இச்சொல் ஆங்கிலத்தில் மொழிப்பெயர்க்கப்படும் போது இது ஒரே சொல்லாக மொழி பெயர்க்கப்படாமல் கடிச வாந ளயமந அடிஇநல டிஇடல என மொழி மொழிபெயர்க்கப்பட வேண்டும்.

2.8.2.2 சொல் வகைப்பாட்டு அடையாளப்படுத்தல் (Parts of Speech Tagging)

சொல் வகைப்பாட்டு அடையாளப்படுத்தி வாக்கியத்திற்கும், சொற்களுக்கும் சொல்வகைப்பாடு தரவேண்டும். இது வாக்கியத்தின் அமைப்பைப் புரிந்துக்கொள்ளவும் வாக்கியத்தின் அமைப்பை உருவாக்கவும் முக்கியமானதாகும்.

தமிழ் வாக்கியம் : அவன் படி ஏறினான்

He staircase Climbed

மேற்கண்ட வாக்கியம் பின்வருமாறு அடையாளப்படுத்தப்படும்.

அவன் < PRO >

படி < N+PRO >

ஏறினான் < V+PST+3SM >

இதில்,

PRO ⇒ Pronoun ‘மாற்றுப் பெயர் ’

N ⇒ Noun ‘ பெயர் ’

NOM ⇒ Nominative ‘ எழுவாய்ப் பெயர் ’

V ⇒ Verb ‘ வினை ’

PST ⇒ Past tense ‘ இறந்த காலம் ’

3SM ⇒ 3 Person Singular, Masculine ‘படர்க்கை ஆண்பால் ’

இந்த வேர்ச்சொல்லான ‘படி’ ‘staircase’ என்ற பொருளையோ ‘read’ என்ற பொருளையோ தரலாம். சோல்வாகைப் பாட்டு அடையாளப்படுத்தி சொல் மற்றும் தகவலைப் பயன்படுத்தி இந்த பொருள் மயக்கத்தைத் தீர்க்கும். இவ்வாறு ‘படி’ என்பதற்குப் பெயர் சொல்வகைப் பாட்டு அடையாளம் தரப்படும்.

2.8.2.3 தொடரியல் பகுப்பாய்வு (Syntax analysis)

தொடரியல் ஆய்வின் போது பகுத்தாப்படுகையில் வாக்கிய அமைப்பு தெரிந்து கொள்ளப்படும். பகுத்தாய்வதால் எச்சத்தொடர் எல்லைகள், முன்னுருபு இணைப்பு அடைகள் என்பன தெரிந்துக்கொள்ளப்பட்டு அவற்றின் பொருத்தமான உறுப்புகளுடன் தொடர்புப் படுத்தப்படும். தமிழ் ஒரு சொல் சுதந்திரமான மொழியாகும். இது தொடரியலில் பகுத்தாய்தலைச் சிக்கலாகும். பெயருடன் இணைந்து வரும் வேற்றுமை உருபுகள் வாக்கியத்தில் இவ்வுறுப்புகளின் பங்களிப்பைக் கண்டு பிடிக்கின்றன. ஆனால் வேற்றுமை உருபு இல்லாது இருப்பு அமைப்பு சார் பொருள் மயக்கத்திற்கு வழிவகுக்கும். எடுத்துக்காட்டாக,

தமிழ் வாக்கியம் : ‘அவள் அக்காள் வீட்டிற்குச் சென்றாள்’

‘ She / he Sister house – dative go+pst 2sf’

என்பதற்கு இரண்டு பகுத்துக் குறிப்புகள் கிடைக்கும் :

(அவள்) NP (அக்காள் வீட்டிற்கு) NP சென்றாள் V

(அவள் அக்காள்) NP (vii TT : RKU) NP சென்றாள் V

2.8.2.4 பொருண்மையியல் பகுப்பாய்வு (Semantic Analysis)

பல அர்த்தங்கள் உள்ள சொற்களுக்கு ஒரு குறிப்பிட்ட சூழலுக்குத் தகுந்தவாறு சரியான அர்த்தம் கண்டுபிடிக்கப்பட வேண்டும். இலக்குமொழி நிகரன் மயக்கம் நீக்கப்பட அர்த்தம் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும்.

தமிழ் வாக்கியம் 1 : அவன் படி ஏறினான்

Climb + pst + 3sm

தமிழ் வாக்கியம் 2 : அவன் படியால் அளந்தான்

He Vessltins Measuvet Ps ++3sm

மேற்கூறிய வாக்கியங்களில் படி என்பதற்கு இரண்டு அர்த்தங்கள் இருக்கின்றன.

1. Staircase
2. A vessel used to measure rice and whet.

முதல் வாக்கியத்தில் ஏறினான் என்ற சொல்லுடன் இணைந்து வருகையைப் பயன்படுத்தி படி என்பதன் அர்த்தம் ‘Staircase’ என்பதுதான் என்றும் இரண்டாம் வாக்கியத்தில் அளந்தான் என்ற சொல்லுடன் சேர்ந்து வருகையால் படி என்பதன் அர்த்தம் ‘vessel’ என்பதுதான் என்றும் நிர்ணயிக்கப்படுகின்றது.

2.8.2.5 தொடரியல் சார் மாற்றம் (Syntactic Transfer)

மொழிபெயர்ப்பு செல்பாட்டில் மாற்றல் அணுகுமுறை (transfer approach) பயன்படுத்தப்படும் போது தொடரியல் மாற்றப்பகுதி தேவைப்படும். ஊள்ளீடு

செய்யப்படும் வாக்கியத்தின் அமைப்பு இலக்குமொழியின் பொருத்தமான அமைப்பிற்குத் தகுந்தவாறு மாற்றப்படுகிறது.

தமிழ் வாக்கியம் : ராமன் புத்தகம் படித்தான்

Raman book read

அமைப்பு: Noun Noun Verb

மேற்கண்ட வாக்கியங்களுக்கு நிகரான வாக்கியங்கள் கீழ்வருவன.

ஆங்கில வாக்கியம் : Raman read a book

அமைப்பு :Noun Verb Noun

இந்த மொழிபெயர்ப்பில் Noun Noun Verb என்ற தமிழ் வாக்கியம் Noun Verb Noun என்று ஆங்கிலத்திற்கு நிகராக மாற்றப்படுகிறது. இந்த எடுத்துக்காட்டு மிக எளிமையானதாகும். ஆனால் இச்செயல்பாடு கலவை வாக்கியங்களுக்கு கிடையில் கடினமானதாகும்.

2 .8.2.6 அகராதி பொருத்தம் (Dictionary Mapping)

இருமொழி அகராதியைப் பயன்படுத்தி மூலமொழி வேர் சொற்கள் இலக்குமொழியில் அதற்கு நிகரான சொற்களுடன் பொருத்தம் செய்யப்படும். ஒரு மூலமொழி வேர்ச்சொல்லுக்கு இலக்கு மொழியில் பல சொற்கள் கொண்ட தொடர்கள் நிகரன்களாக வரலாம். சூழல் அடிப்படையில் பொருத்தமான சொல் சூழல் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும். வேர்ச் சொல்லின் இலக்கண பண்புக் கூறுகளின் இலக்கணம் அகராதியில் அடங்கி இருக்கும்.

2.8.2.7 உருபனியல் உருவாக்கம் (Morphdogical Generation)

இங்கு சொல்லின் திரிபு வடிவம் வேர் மற்றும் அதன் இலக்கண பண்புக் கூறுகளால் உருவாக்கப்படும்.

தமிழ் வேர்ச்சொல் : 'படி'
இலக்கண பண்புக் கூறுகள் : 'Past tense'+ 3rd person, Singular, Mascuine'

சொல் வடிவு : ' படித்தான் '

2.8.3 தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்

2.8.3.1 கருவிகள் (Tools)

2.8.3.1.1 உருபனியல் பகுப்பாய்வி (Morphoogical Analyzer)

95 விழுக்காட்டிற்கு மேல் முழுமையுடன் தொழில் நுட்பத்தின் தற்போதைய நிலையைப் பயன்படுத்தி ஒரு திறமையான உருபனியல் ஆய்வி தேவை. தற்போதைய உருபனியல் ஆய்விகள் மேற்கோள்வாய்வாய்பாட்டு (Paradim approach) அணுகுமுறையையும் முற்றுநிலைத் தானியங்கியையும் (Finite State Automata) பயன்படுத்துகின்றனர்.

2.8.3.1.2 அடையாளப்படுத்தி

வாக்கியங்களை அடையாளப்படுத்த முழுதானியக்க, பகுதிதானியக்க, விதி அடிப்படையிலான அல்லது புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தித் தேவை. இவ்வகையான அடையாளப்படுத்தல் இயந்திரமொழி பெயர்ப்புக்கு மட்டுமல்லாமல் தகவல் பிரித்தெடுப்பு, தகவல் மீட்பு உரை சுருக்கம் மற்றும் பிற பல்வேறுபட்ட பயன்பாடுகளிலும் பயன்படுகிறது.

2.8.3.1.3, பகுப்பான் (Parser)

ஒரு பகுப்பான் சார்பு இலக்கணம் (dependency Gramman) தொடரமைப்பு இலக்கணம் (Phrase Structure Grammer), கிளை இணைப்பு இலக்கணம் (Tree Adjoining Grammar) போன்ற ஏதாவது இலக்கண வடிவமைப்புகள் அடிப்படையில் வாக்கியங்களுக்குப் பகுப்புக் கிளை தரும். ஒரு பகுப்பான் தமிழ்மொழிப் பகுப்பாய்விற்கு மிகத் தேவை. இது விதிகள் மற்றும் பகுக்கும் இயந்திரத்தை உருவாக்குதலையும் உள்ளடக்கும். சில இயற்கை மொழி ஆய்வுப் பயன்பாடுகள் முழு பகுத்தலை வேண்டாது, அவைகளுக்குப் பகுதி பகுத்தல் போதுமானது. பெயர்த்தொடர் கண்டுபிடிப்பு, எச்சத்தொடர் கண்டுபிடிப்பு என்பன இந்தப் பயன்பாடுகளுக்கு உதவும்.

2.8.3.1.5 பொருண்மையியல் பகுப்பாய்வி (Semantic Analyser)

சொல் அர்த்த மயக்கம் நீக்கும் கருவியை உருவாக்கும் வேலை அதன் தொடக்க நிலையில் தான் இருக்கிறது. இந்த வேலை கூடுதலான கவனக் குவிப்பையும் கடினமான முயற்சிகளையும் வேண்டும். மேலும் இது சொற்கள் இணைந்து வரும் தகவலைச் சேகரிக்க வேண்டி பெரிய அடையாளப்படுத்தப்பட்ட விரிதரவை வேண்டும்.

2.8.3.1.6 உருபனியல் உருவாக்கி (Morphological Generator)

நம்மிடம் உருபனியல் ஆய்வுக்கு நன்றாக அமைக்கப்பட்ட முற்றுநிலை இயந்திரம் இருந்தால் அதை உருபனியல் உருவாக்குதலுக்குச் சிறிது மாற்றமின்ற அல்லது சிறிது மாற்றத்துடன் அதே இயந்திரத்தைப் பயன்படுத்தலாம். இவ்வாறு

உருவாக்கப்பட்ட உருபனியல் உருவாக்கிகள் திருப்தியான விளைவுகளைக் காட்டுகின்றன.

2.8.3.2 மூலவளங்கள் அகராதி (Resources)

2.8.3.2.1 அகராதி (Dictionary)

பல அகராதிகள் தமிழில் இருந்தாலும் அவையெல்லாம் மின்வடிவில் இல்லை. இந்த அகராதிகள் இயந்திரம் படிக்கவியலும் வடிவத்தில் இருந்தால்தான் அவற்றைக் கணினி செயல்பாட்டிற்குப் பயன்படுத்த இயலும்.

2.8.3.2.2 மொழி கடந்த அகராதி (Cross -lingual Dictionary)

இது சாதாரண மொழிகள் கடந்த அகராதியைக் குறிப்பிடவில்லை. இயந்திர மொழிபெயர்ப்புக்கு நமக்கு தேவையான அகராதி மொழி கடந்த வேர் அகராதிகளாகும். தனிப்பட்ட மனிதர்களும் நிறுவனங்களும் அவர்களுடைய பயன்பாட்டிற்காக இருமொழி அகராதிகளை உருவாக்குகின்றனர். இவை இந்த ஆய்வு களத்தில் சூடுபடும் ஆய்வாரள்களுக்கிடையில் பங்கிடப்படவேண்டும்.

2.8.3.2.3 தரவுத்தொகுதி (Corpus)

ஒரு மொழி தரவுத்தொகுதி அம்மொழியின் நடை, சொல், பயன்பாடு போன்றவற்றின் உருப்படுத்தம் ஆகும். இது மேலும் உருபனியல் பகுப்பாவிகள், பகுப்பான்கள் மற்றும் உருவாக்கிகள் என்பனவற்றை மதிப்பீடு செய்யும் பரிசோதனைக் கருவியாகச் செயல்படும். இந்திய மொழிகளின் மைய நிறுவனம் சுமார் 15 மொழிகளுக்கு மூன்று மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதிகளை உருவாக்கியுள்ளது. மேலும் சில தரவுத்தொகுதிகளை

அடையாளப்படுத்தி உள்ளது. இருக்கின்ற வசதிகளை வைத்துக்கொண்டு பல்வேறுபட்ட மொழிகளுக்கு விரிதரவைச் சேகரிக்க வேண்டி பொதுவான திட்டம் ஊக்கப்படுத்தப்பட வேண்டும். புத்தகங்கள், நாளிதழ்கள், கால இதழ்கள் போன்றவற்றஇ தவிர இணையதளங்களைப் பயன்படுத்தி தரவுத்தொகுதி சேகரிக்கப்பட்ட வேண்டும்.

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட விரிதரவில் (POS Tagged Corpus) உள்ள எல்லா வாக்கியங்களும் சொல்வகைப்பாட்டிற்கு வேண்டி அடையாளப்படுத்தப்பட்டிருக்க வேண்டும். இந்த அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, புள்ளியியல் அடிப்படையிலான சொல் வகைப்பாட்டு அடையாளப்படுத்தி உருவாக்குவதற்குத் துணைபுரியும். இம்மூலவளத்தின் மீது பல்வேறு வகையான ஆய்வுகளினால் ஒரு மொழிக்குக் கணினி சார் இலக்கணம் உருவாக்க இயலும்.

இணை தரவுத்தொகுதிகள் (Parellel Corpora)

இணைதரவுத்தொகுதி இருமொழிகளுக்கு மாற்றமைவு விதிகளை (transfer rules) உருவாக்க முக்கியமான மூலவளமாகும். இது புள்ளியியல் அல்லது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறைகள் உருவாக்கத் துணைபுரியும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கு முன் வரிசையாக்க விதிகள் இணைவிரிதரவிகளின் பத்திகள், வாக்கியங்கள் சொற்கள் இவற்றை வரிசையாக்கம் செய்ய வரிசையாக்கக் கருவிகள் தேவைகள். சுப்ல வெளிப்யீட்டாளர்களிடமும் அரசு நிறுவனங்களிலும் இம்மாதிரியான இணைதரவுத்தொகுதிகள் இருக்கின்றன. இந்த

இணைதரவுத்தொகுதிகள் விளைவாக்கம் உள்ள ஒழுங்குமுறைகளை உருவாக்க வேண்டி ஆராய்ச்சியாளர்களுக்குப் பயன்படுத்தத் தரப்படவேண்டும்.

2.8.3.2.4 மாற்றமைவு இலக்கணம்

மாற்றமைவு இலக்கணம் (transfer grammar)இருமொழிகளின் தொடரியல் அமைப்புக்கு இடையில் உள்ள முரண்வேறுபாடுகளை அறிந்து கொள்கின்றது. மாற்றமைவு அணுகுமுறைகளை உருவாக்க இம்மாதிரியான மாற்றமைவுப் பகுதியை உருவாக்குவது கட்டாயமாகும். மொழிகளைப் பற்றிய ஆழ்ந்த அறிவுள்ள மொழியியலார் இம்மாதிரியான மூல வளத்தை உருவாக்கலாம்.

2.8.3.2.5 இணைவமைதி அகராதி

இணைவமைதி அகராதி (Collocation dictionary) மொழியின் மயக்கமான சொற்களின் முறையான சேர்க்கைகளைக் கொண்டிருக்கும். அம்மாதிரியான அகராதி அர்த்த மயக்க நீக்கிற்கு முக்கியமான பகுதியாகும். தமிழுக்கு அம்மாதிரியான மூலவளத்தைச் சில ஆய்வாளர் குழுமங்கள் செய்து வருகின்றன.

2.9 நிறைவுரை

இவ்வியலில் கணிப்பொறி வழி இயந்திர மொழிபெயர்ப்பின் தேவை, கணீபொறி வழி மொழிபெயர்ப்பின் வகைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும், இந்தியாவில் இயந்திர மொழி பெயர்ப்பின் வளர்ச்சி, தமிழ்ச்சார்ந்த இயந்திர மொழிபெயர்ப்புகள் என்ற தலைப்புகளில் விளக்கங்கள் தரப்பட்டுள்ளன.

இயல் 3

தமிழுக்குத் தரவுத் தொகுதிகளை உருவாக்குதல்

3.1 அறிமுகவுரை

அனுபவவாத ஆய்வுகள் எந்த எழுதப்பட்ட அல்லது பேசப்பட்ட உரைகளையும் பயன்படுத்தி அதன் மேல் செய்யப்படுவதாகும். இம்மாதிரியான தனிப்பட்ட உரைகள் பலவகையான இலக்கிய மற்றும் மொழியியல் ஆய்வுகளுக்கு அடிப்படையாக அமைகின்றது. எடுத்துக்காட்டாக, ஒரு செய்யுளின் அல்லது ஒரு புதினத்தின் நடையியல் ஆய்வு அல்லது ஒரு தொலைக்காட்சிப் பேச்சின் உரையாடல் ஆய்வு என்பனவற்றைக் கூறலாம். ஆனால் அனுபவவாத மொழியியலின் ஒரு வடிவுக்கு அடிப்படையாக அமையும். ஒரு குறிப்பிட்ட பகுதியைப் பரிசோதிக்கும் வழிகளிலிருந்து பல அடிப்படையான வழிகளில் வேறுபடும். கொள்கை அடிப்படையில் ஒரு உரைக்குக் கூடுதலான எந்தச் சேகரிப்பு தொகுதியும் தரவுத் தொகுதி அல்லது தரவுத்தொகுதி எனப்படும். தரவுத்தொகுதி என்று தமிழில் அழைக்கப்படும். corpus என்பது Body என்பதன் இலத்தீன் சொல்லாகும். எனவே தரவுத்தொகுதி என்பதை எந்த ஒரு உரையின் உடல் என்று கூறலாம். ஆனால் தரவுத்தொகுதி என்ற சொல் தற்கால மொழியியல் என்ற சூழலில் பயன்படுத்தப்படும் போது எளிய விளக்க உரையைத் தாண்டி கூடுதல் சிறப்பு அர்த்தங்களைக் கொண்டிருக்கும் போக்கு விரிதரவில் காணப்படும்.

தரவுத்தொகுதியின் இரண்டு முக்கிய பகுதிகள்

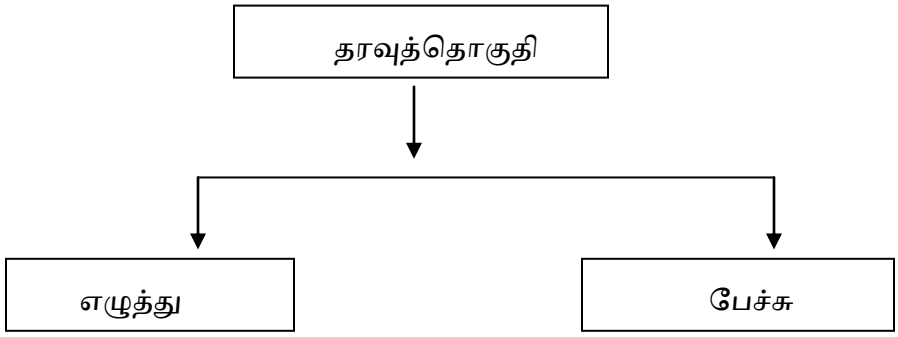
1. தகவல் உரைநடையின் உரைகள்
2. கற்பனை உரைநடையின் உரைகள்

தகவல் உரைகள் பின்வருவனவற்றை உள்ளடக்குகின்றது:

- பத்திரிக்கை : அறிக்கை (அரசியல், விளையாட்டு, சமுதாயம், வட்டாரச் செய்திகள், பொருளாதாரம், பண்பாடு)
- பத்திரிக்கை : இதழாசிரியர் கருத்து (நிறுவனம் சார்ந்தவை, பதிப்பாசிரியருக்கு தனிநபர் கடிதங்கள்)
- பத்திரிக்கை : திறனாய்வு (சினிமா, புத்தகம், இசை, நடனம் போன்றவை)
- மதம் (புத்தகங்கள், கால இதழ்கள், சமயக்கட்டுரைகள்)
- திறமைகளும், பொது போக்குகளும் (புத்தகங்கள் கால இதழ்கள்)
- புகழ் வாய்ந்த மரபுச் செய்திகள் (புத்தகங்கள், கால இதழ்கள்)
- கடிதங்கள், வாழ்க்கை வரலாறு, நினைவுகள்
- பலதரப்பட்டவை (அரசாங்க ஆவணங்கள், அடித்தள அறிக்கைகள் தொழில் நிறுவன அறிக்கைகள், கல்லூரிப் பெயர்ப்பட்டியல் தொழில் நிறுவனம், வீடு)
- கற்றல் மற்றும் அறிவியல் எழுத்துரைகள் (இயற்கை அறிவியல் மருத்துவம், கணினித் தொழில் நுட்பம், பொறியியல்)
- சமூக மற்றும் நடத்தை அறிவியல்கள் (அரசியல் அறிவியல், சட்டம் கல்வி) மனிதவியலிலிருந்து உரைகள்

கற்பனை உரைநடையின் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

- பொதுக் கதைகள் (புதினம், சிறுகதைகள்)
- அறிவியல் கதைகள் (புதினங்கள் , சிறுகதைகள்)
- துணிக்கரசி செயல் பற்றிய மற்றும் மேற்கத்திய கதைகள் (புதினங்கள் , சிறுகதைகள்)
- காதல் கதைகள் (புதினங்கள் , சிறுகதைகள்)
- நகைச்சுவை (புதினங்கள் , சிறுகதைகள்)



3.2 ஆங்கில மொழிக்கான தரவுத்தொகுதிகள்/தரவுத்தொகுதிகள்

3.2.1 பிரவுன் தரவுத்தொகுதி

முதல் மின் தரவுத்தொகுதி 1961-ல் பிரவுன் பல்கலைக்கழகத்தில் நெல்சன் பிரான்சிஸ் ஷென்றி குச்சேரா (W.Nelson francies and Henry kuchera) என்பவர்களால் உருவாக்கப்பட்ட பிரவுண் விரிதரவாகும் (Brown corpus (the standard sample of present day American English). இது 1961-ஆம் ஆண்டு நாள் காட்டி ஆண்டு யுஎஸே-ல் USA அச்சடிக்கப்பட்ட ஆங்கில உரை நடையின்ன தொடர்ச்சியான பனுவலின் 1014312 சொற்களைக்கொண்டது. இது 2000ரூ சொற்களைக் கொண்ட 500 மாதிரிகளால் (Sample) பகுக்கப்பட்டள்ளது. ஒவ்வொரு மாதிரியும் ஒரு வாக்கியத்தின் தொடக்கத்தில் தொடங்குகிறது. ஆனால் ஒரு பத்தியின் அல்லது ஒரு பெரிய பகுதியின் தொடக்கத்தில் தொடங்குகின்றது என்பது தேவை இல்லை. ஒவ்வொன்றும் 2000 சொற்களுக்குபின் இறுதியும் முதல் வாக்கியத்தில் முடிகிறது. ஒவ்வொரு மாதிரியும் உரைநடையின் நடைகள் மற்றும் வகைகளின் விரிந்த பரப்பு எல்லையைப் பிரதி நிதித்துவம் செய்யும். செய்யுள்களும், நாடகங்களில் உள்ள உரைகளும் எழுத்துக் கருத்தாடல்களுக்கு அப்பாற்பட்ட பேச்சு கருத்தாடலின் கற்பனையான பொழுது போக்காக இருப்பதன் காரணமாக அவைகள் விலக்கி வைக்கப்படுகிறது. கதைகள்

உட்படுத்தப்பட்டுள்ளன. ஆனால் 50 விழுக்காடுக்கு மேல் உரையாடல் உள்ள மாதிரிகள் எடுத்துக்கொள்ளப்படமாட்டாது. முஹ்திரிகள் அவற்றின் பிரதிகளித்துவப் பின்பிற்கேற்ப வேறு எந்த அகவயமாகத் தீர்மானிக்கப்பட்ட சிறப்புக்காகவும் தேர்ந்தெடுக்கப்படவில்லை.

3.2.2 லோப் தரவுத்தொகுதி

லங்காஸ்டர் ஒசலோ பெர்கன் தரவுத்தொகுதி (Lancaster Oslo / Bergen (LOB)) என்பது பிரிட்டிஷ் ஆங்கிலத்தின் ஒரு மில்லியன் சொல் சேகரிப்பாகும். இது லங்காஸ்டர் பல்கலைக்கழகத்தைச் சார்ந்த ஜியாபிரே லீச் (Geofferey leech) மற்றும் ஒசலோ பல்கலைக் கழகத்தைச் சார்ந்த ஸ்டிக் ஜோகன்சன் (Slig Johansson) என்போர் பெர்கனில் உள்ள மனிதவியலுக்கான நார்வேஜியன் கணிப்பு மைத்துடன் இணைந்து தொகுக்கப்பட்டதாகும். லோப் யுகே–இல் (UK) 1961-ஆம் வருடத்திற்குள் வெளிப்யிடப்பட்ட மூலப்பெருட்களைக் கொண்டது. இது பிரவுன் விரிதரவைப் போல் உரை மாதிரிகளைக் கொண்டது. லோப் தரவுத்தொகுதியின் திட்ட நெறிமுறை யுஎஸே இலும், யுகே-விலும் பயன்படுத்தப்படும் ஆங்கிலத்தின் இரு வகைகளுக்கிடையே உள்ள எதிர்கால ஒப்பீட்டிற்கு வேண்டி பிரவுன் தரவுத்தொகுதிடன் ஒற்றுமைக் கொண்டுள்ளது.

3.2.3 ஆங்கிலத்தில் ஆஸ்டிரேலியன் தரவுத்தொகுதி

ஆங்கிலத்தில் ஆஸ்டிரேலிய தரவுத்தொகுதி (Australian corpus of English (ACE)) என்பது ஆஸ்டிரேலியாவிலுள்ள மாக்குயர் பல்கலைக் கழகத்தில் (Macquaire University) மொழியியல் துறையில் 1986-1990 என்ற கால கட்டத்தில் தொகுக்கப்பட்டது. ஏசியி 1986க்குள் வெளிப்யிடப்பட்டது மூலப் பொருட்களைக் கொண்டிருக்கிறது. ஏசியி பல்வேறுபட்ட மொழியியல் ஆய்வுகளில் உதவுவதற்காகத் திட்டமிடப்பட்ட ஆஸ்டிரேலியாவில் முதலில் தொகுக்கப்பட்ட

பலபடித்தான விரிதரவாகும். இது தற்கால ஆஸ்டிரேலிய ஆங்கிலத்தின் திறனான மாதிரியாகும். இது ஆஸ்டிரேலியாவில் கூடுதல் சிறப்புப் பண்புள்ள ஒரு படித்தான தரவுத்தொகுதிகளுடன் ஒப்பீட்டுப் பார்ப்பதற்கு வேண்டி நோக்கிட்டு விரிதரவாகச் செய்யப்படுகிறன்றது. இது பிரவுன் விரிதரவிலும், லோப் விரிதரவிலும் பிரிதி நிதித்துவம் தரப்பட்டுள்ள இனங்களின் சரிசமத்திற்குப் பொருத்துவத நோக்கமாகக் கொண்டது. எனவே இது ஏறக்குறைய ஒவ்வொரு வகைப்பாட்டிலும் 2000 சொல் மாதிரிகளின் நிகரான குழுமத்தை உருவாக்குகிறது. துணை இனங்களையும், விக்ஷயப்பரப்புகளையும் பொறுத்துவது தேவை என்பதால் தரவுத்தொகுதி வகைப்பட்டிற்குள்ளும் மாதிரிச் செயல் முறைகள் பெரும்பாலும் திட்டமிடப் பட்டுள்ளது. (வரைமுறையற்றதல்ல)

3.2.4 எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி

எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி (Wellington Written New Zealand English) என்பது 1982 கால கட்டத்தில் வெல்லிங்டன் விக்டோரியா பல்கலைக் கழகத்தில் உருவாக்கப்பட்டது. நியூசிலாந்து ஆங்கில தரவுத்தொகுதியின் நோக்கம், பிரவுன் விரதரவு, லோப் தரவுத்தொகுதி ஏசியி தரவுத்தொகுதி என்பதுடன் நேரடியாக ஒப்பீடு செய்வதை அனுமதிக்கும் எழுத்து நியூசிலாந்து ஆங்கிலத்தின் கணினியாக்கம் செய்யப்பட்ட மாதிரியைத் தருகிறது. முக்கிய உரை வகைப்பாடுகள் எவ்வளவு இயலுமோ அவ்வளவு லோப் தரவுத்தொகுதிடன் பொருந்தும்படி வரிசைப் படுத்தப்பட்டுள்ளது. கற்பனை உரைநடையில் குறிப்பாக கதைகளில் தேவைகருதிச் சில மாற்றங்கள் செய்யப்பட்டுள்ளன.

இந்திய ஆங்கிலத்தில் கோலப்பூர தரவுத்தொகுதி (Kollapur Corpus of India)எஸ்-வி சாஸ்திரியும். அவரைச் சார்ந்தவர்களும் இணைந்து கோலாப்பூரில் உள்ள சிவாஜி பல்கலைக்கழகத்தில் 1988-ஆம் ஆண்டில் வெளிப்யிடப்பட்ட

விசயங்களிலிருந்து எடுக்கப்பட்ட இந்திய ஆங்கிலத்தின் ஒப்பீட்டு ஆய்வுகளுக்கு வேண்டிய மூல சாதனங்களாகப் பயன்படுகின்றது. இது இந்திய ஆங்கிலத்தின் விரிந்த வர்ணனைக்குக் கொண்டுச் செல்லும் என்று எதிர்பார்க்கப்படுகிறது. லோம் மற்றும் பிரவுன் தரவுத்தொகுதிகளை உருவாக்க பயன்படுத்த திட்டத்தைப் பின்பற்றி கோலாப்பூர் தரவுத்தொகுதிக்கு 15 விசுஷய வகைப்பாடுகளிலிருந்து உரைகள் தேர்ந்தெடுக்கப்பட்டுள்ளன.

3.2.5 லோப் தரவுத்தொகுதி

ஃப்லோப் (ஃப்ரெல்பர்க் லோப் தரவுத்தொகுதி (Freiburg lob Corpus for British English (Flob)) மற்றும் லோப் தரவுத்தொகுதிகளுடன் பொருத்தும் ஒரு குழும தரவுத்தொகுதிகளைத் தொகுக்கும் முயற்சியின் விளைவாகும். இது 1990-களின் தொடக்க நிலை மொழியைப் பிரதிநிதித்துவம் செய்யும் என்ற நிலையில் வேறுபடும் இத்திடம் 1990-ஏப்ரலில் ஜெர்மனியில் கிருஷ்டியன் மேரியின் கீழ் தொடங்கப்பட்டது.

இது மொழியாலாளர்களுக்குப் பின் வரும் செயல்களில் உதவுகிறது.

- இன்றைய ஆங்கிலத்தில் மொழிசார் மாற்றத்தின் மீதான இன்றைய கருதுகோளாப் பரிசோதிக்க.
- சொல் நிகழ்வெண்களின் முறையான ஒப்பீடுகளின் வழி இலக்கியத்தில் இதுவரை கண்டுப்பிடிக்கப்படாத மாற்றங்களைக் கண்டுபிடிக்க
- ஒரு கால இட வழக்குகளின் (British and America) தொடர்ச்சியான மாற்றம் மற்றும் நடையியல் வேறுபாடு இவற்றிற்கு இடையிலான சார்பையும் மெய்யான இருகால வளர்ச்சியையும் ஆயும் முக்கியமான நெறிமுறைச் சிக்கல்களில் ஒன்றை முறையாகக் கையாளுவது. இப்புதிய பிரிட்டிஷ் மற்றும் அமெரிக்க தரவுத்தொகுதிகளின் கூடுதல் பயன்

என்னவென்றால் அவை இந்திய, ஆஸ்திரேலிய மற்றும் நியூசிலாந்து தரவுத்தொகுதிகளுடன் (பிந்தைய 1980-களில் உளடள மொழி பயன்பாட்டை பிரதிநிதித்துவம் செய்யும் மாதிரிகள்) ஒப்பிடமூல லோப் மற்றும் பிரவுன் தரவுத்தொகுதிகளைவிடக் கூடுதல் பொருத்தமான தரவு மையங்களைத் தருகின்றன.

3.2.6 பிரிட்டிஷ் தேசிய தரவுத்தொகுதி

பிரிட்டிஷ் தேசிய தரவுத்தொகுதி (British National Corpus (BNC) என்பது மூலங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெறப்பட்டு எழுதப்பட்ட மற்றும் (10%) பல வயதினர்கள், வட்டாரங்கள், சமூக வகுப்புகள் இவற்றிலிருந்து பெறப்பட்ட எழுதப்படாத முறைசாரா உரையாடல்கள் மற்றும் முறையான வணிக மற்றும் அரசாங்கக் கூட்டங்களிலிருந்து வானொலி மற்றும் தொலைபேசி வரையிலான பரப்பெல்லையைக் கொண்ட சூழல்களின் எல்லா வகைகளிலிருந்தும் சேகரிக்கப்பட்ட பேச்சு மாதிரிகள் என்பனவற்றை உள்ளடக்கும்.

இப்பொதுத்தன்மை பிஎன்சி விரிதரவை அகராதியியல், பொருண்மையியல், உருபனியல், மொழிப் பகுப்பாய்வு, செயற்கை அறிவுநுட்பம், பேச்சு அறிதல் மற்றும் கூட்டிணைப்பு, இலக்கிய ஆய்வுகள் மற்றும் முக்கிய மொழியியலின் எல்லாப் பரப்புகளையும் உள்ளடக்கும் ஆய்வு நோக்கங்களின் விரிந்த வகைக்குப் பயனுள்ளதாகச் செய்கின்றது.

பிஎன்சி பேச்சு உரையாடல் மற்றும் ஒருவர் உரையாடல் இவற்றிலிந்து ஒலிபெயர்ப்பு செய்யப்பட்ட 863 உரைகளை உள்ளடக்கிய 4124 உரைகளைக் கொண்ட 100106008 சொற்களை உள்ளடக்கியது. ஒவ்வொரு உரையும் மேலும் எழுத்துவடிவ வாக்கிய அலகுகளாகப் பிரிக்கப்பட்டுள்ளது. இதற்குள் ஒவ்வொரு சொல்லுக்கும் சொல் வகுப்பு (சொல்வகைப்பாடு) குறியம் (code) தரப்பட்டுள்ளது.

பிரித்தலும் சொல் வகைப்படுத்தலும் (claws) சொல்வகைப்பாட்டு அடையாளப்படுத்தியால் தானியக்கமாகச் செய்யப்படுகின்றது. பிஎன்சி விரிதரவில் பயன்படுத்தப்படும் வகைப்பாட்டுத் திட்டம் 65 சொல் வகைப்படுகளை வேறுபாடுகளை வேறுபடுத்துகின்றது.

3.2.7 அமெரிக்க தேசிய தரவுத்தொகுதி

அமெரிக்க தேசிய தரவுத்தொகுதி (American National Corpus (ANC)) பிஎன்ச தரவுத்தொகுதி ஒப்பிடத்தக்க அமெரிக்க ஆங்கிலத்தை உட்படுத்திய தரவுத்தொகுதியின் உருவாக்கத்திற்கு உதவியது. இதன் நோக்கம் பிஎன்சி தரவுத்தொகுதிப் பொதுவினங்களுடன் ஒப்பிடத்தக்கக் குறைந்தது 100 மில்லியன் சொற்களைப் பெறுவதாகும். இது மொழித் தொழில் நுட்ப ஆய்வுக்கு உதவுவதுடன் எல்லா மட்டங்களிலும் மொழிப் பகுத்தாய்வுக்கும் கல்விக்கும் பயன்படவேண்டி செழுமையான தேசிய மூலத்தைத் தருகின்றது. எஎன்சி அமெரிக்க ஆங்கில அகராதிகளின் வெளிப்யீட்டார்கள் மற்றும் மொழி ஆய்வில் ஆர்வமுள்ள நிறுவனங்கள் இவற்றின் கூட்டிணைப்பின் (consortium) பங்களிப்பு வழி உருவாக்கப்பட்டது. மொழித் தரவுக் கூட்டிணைப்பும் (linguistics Data consortium LDC) சாதன மற்றும் பொருளாதார ஆதரவும் தந்தது. ஏஎன்சி தரவுத்தொகுதியின் 10 மில்லியன் சொற்களின் முதல் வெளிப்யீட்டில் உள்ளடக்கப்பட்ட உரைகள் முதலில் பெறப்பட்டவையாகும். எனவே, தரவுத்தொகுதி சரிநிகரானதல்ல பேச்சுப் பகுதி 3224388 சொற்களையும் எழுத்துப்பகுதி 8283828 சொற்களையும் கொண்டுள்ளது.

3.2.8 ஆங்கில வங்கி

ஆங்கில மொழி வங்கி (Bank of English) சொற்கள், இலக்கணம் மற்றும் பயன்பாடு (usage) என்பவனவற்றின் பகுப்பாவிற்கு வேண்டி வடிவமைக்கப்பட்ட

தற்கால ஆங்கில மொழியின் மாதிரிகளின் சேகரிப்பாகும். ஆங்கில மொழி வங்கி 1991-ன் (Cobuild (Harper Collins Publishers))மற்றும் யுகேயில் உள்ள பிர்மிங்காம் பல்கலைக்கழகம் இவற்றால் 1991-ல் வெளிப்யிடப்பட்டது. இது புதிய சாதனங்களின் நிரந்தர சேர்ப்பினால் தொடர்ந்து வளர்கின்றது. சேகரிப்பு எழுதப்பட்ட மற்றும் பேச்சின் வேறுபட்ட வகைகளின் விரிந்த பரப்பெல்லையைக் கொண்டது. இது நூற்றுக்கணக்கான வேறுபட்ட மூலங்களிலிருந்து ஆங்கில மொழியின் மாதிரிகளைக் கொண்டிருக்கின்றது. எழுதப்பட்ட உரைகள் செய்தித்தாள்கள், இதழ்கள், (magazines) கதைப் புத்தகங்கள் மற்றும் கதையல்லாப் புத்தகங்கள், சிற்றேடுகள் (brpcjers) துண்டுப்பத்திரிக்கைகள் (leaflets) அறிக்கைகள், கடிதங்கள் மற்றும் பிறவற்றிலிருந்து எடுக்கப்பட்டவைகளாகும். பேச்சுச் சொற்கள் அன்றாட சாதாரண உரையாடல்கள், வானொலி ஒலிபரப்புகள், கூட்டங்கள் நேர்காணல்கள் மற்றும் விவாதங்கள் போன்றவற்றால் பிரதிநிதித்துவம் செய்யப்பட்டுள்ளன.

3.3 இந்திய மொழிகளுக்கான தரவுத் தொகுதிகள்

3.3.1 இந்திய மொழிகளின் MIT தரவுத்தொகுதி

இந்திய மொழிகளுக்கு தரவுத்தொகுதி உருவாக்குவது 1991-ல் தொடங்கப்பட்டது. இது இந்திய அரசின் மின்னுத் துறை எல்லா இந்திய மொழிகளுக்கும் இயந்திரத்தால் படிக்கவியலும் உரைகளின் தரவுத்தொகுதிகளஇ உருவாக்க வேண்டி இந்திய மொழிகளுக்குத் தொழில் நுட்ப வளர்ச்சி (Technological Development for Indian Languages (TDIL)) என்ற திட்டத்தைத் தொடங்கியது. இதே நேரத்தில் மொழி ஆய்வுக்கும் (சொல்வகைப்பாடு அடையாளப்படுத்தி, உரை குறியாக்கி, புள்ளியல் எண்ணி, எழுத்துத் திருத்தி, உருபனியல் பகுப்பாய்வி போன்றவற்றை உருவாக்க) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு எந்திர உதவிசார் மொழிபெயர்ப்புக்கு (machine-aided translation) வேண்டி கருவிகளை வடிவமைத்தல் என்ற செயல்பாடுகளுக்கு மென்பொருள்களை

உருவாக்கவும் ஊக்கம் தரப்பட்டது. டெல்கியில் உள்ள இந்தியத தொழில் நுட்ப நிறுவனம் (Indian institute of technology) இந்திய ஆங்கிலம் , இந்தி மற்றும் பஞ்சாபி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் தமிழ், தெலுங்கு, கன்னடா மற்றும் மலையாள மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. பூனேயில் உள் டெக்கான் கல்லூரி மராத்தி மற்றும் குஜராத்தி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. புபனேஸ்வரிலுள்ள பயன்பாட்டு மொழி அறிவியல்களின் இந்திய நிறுவனம் ஒரியா, வங்காளம், அசாமீஸ் ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. வாரணாசியிலுள்ள சம்பூர்ணநந்தா சமஸ்கிருதப் பல்கலைக்கழகம் சமஸ்கிருத மொழிக்கு தரவுத்தொகுதி உருவாக்கியது. அலிகார் முஸ்லீம் பல்கலைக்கழகம் உருது, சிந்தி மற்றும் காஷ்மீரி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது.

கான்பூரிலுள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் மொழி பகுத்தாய்விற்கும் இயந்திர மொழிபெயர்ப்புக்கும் கருவிகளையும் மென்பொருள்களையும் வடிவமைக்கும் பொறுப்பை ஏற்றது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் எல்லா மொழிகளின் முழு தரவுத்தொகுதி தரவுமையத்தைச் சேகரித்துப் பாதுகாக்கும் பொறுப்பை ஏற்றது. 1995-ல் இத்திட்டத்தின் முடிவுற்றபோது மின்னனு நிறுவனம் இது நீண்டகால ஆய்வையும் பெரிய முதலீட்டையும் வேண்டும் என்று உணர்ந்து திட்டத்தைத் தொடராமல் நிறுத்தியது. இருப்பினும் தற்போது இந்திய அரசின் MIT புதிய வெளிப்ச்சத்திலும் பார்வையிலும் இச்செயல்பாட்டை மீண்டும் நடைமுறைப்படுத்தியுள்ளது.

3.3.2 இந்திய மொழிகளுக்கான தரவுத்தொகுதி

இந்திய மொழிகளுக்கான தரவுத்தொகுதி பற்றி Jayaram மற்றும் Rajashree விரிவாக விளக்கியுள்ளனர். மேலும் இந்திய மொழிகளின் நடுவண் நிறுவனம்

இந்தி, பஞ்சாபி, காஷ்மீரி, உருது, கன்னடம், தமிழ், தெலுங்கு, மலையாளம், ஒரியா, அசாமி மற்றும் பெங்காலி போன்ற பன்னிரெண்டு மொழிகளுக்கும் கிட்டத்தட்ட மூன்று மில்லியன் சொற்கள் வருமாறு விரிதரவை உருவாக்கியுள்ளது.

தரவுத்தொகுதியின் நோக்கம்

தரவுத்தொகுதி இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட தரவுத்தொகுதிகள் பல்நோக்கு தரவுத்தொகுதிகள் ஆகும். இந்த ஒரு விரிதரவை உருவாக்க செலவான தொகை மற்றும் தொகுக்க தேவையான நேரம் என்பன பல்வேறுபட்ட மொழியியல் ஆய்வுகளுக்கு இவற்றை அடிப்படை மூலவளமாக பயன்படுத்த ஊக்கம் அளித்து வருகிறது. ஒருவர் தரவுத்தொகுதி செய்ய எல்லா ஆய்வுகளுக்கும் மிகப் பொருத்தமான முன்மாதிரி என்று எடுத்துக்கொள்ள இயலாது. தரவுத்தொகுதியின் நோக்கம் பின்வருமாறு:

1. வேறுபட்ட உரை வகைகளிலிருந்து சொற்கள் மற்றும் வாக்கியங்களின் மாதிரி.
2. வேறுபட்ட உரை வகைகளில் சொற்களின் நிகழ்வெண்
3. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
4. வேறுபட்ட சுழல்களில் சொற்களின் தேர்வு
5. தனிநிலை சொற்கள் சேர்ந்து வருகை மற்றும் இலக்கண – தொடரியல் அமைப்பொழுங்கு
6. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
7. பயன்பாட்டடின் இயற்கையான எடுத்துக்காட்டுகளின் தேர்வுக்குக் கச்சாப்பொருள்.

3.3.3 EMILLE – தரவுத்தொகுதி

EMILLE [Enabling Minority Language Engineering] தரவுத்தொகுதிகள், Emille திட்டம், யு.கே.யில் உள்ள லங்காஸ்டர் பல்கலைக்கழகம், மைசூரில் உள்ள இந்திய மைய நிறுவனம் இவற்றின் ஒத்துழைப்பால் உருவாக்கப்பட்டது. Emille ஐரோப்பிய மொழி மூலவளங்கள் சங்கத்தால் [European Language Resources Association] வழங்கப்பட்டது. இந்த தரவுத்தொகுதி மூன்று பகுதிகளைக் கொண்டது.

1. ஒரு மொழி தரவுத்தொகுதி [Monologual Corpus]
2. இணை தரவுத்தொகுதி [Parallel Corpus]
3. அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி [Annotated Corpus]

ஆசாமி, பெங்காளி, குஜராத்தி, இந்தி, கன்னடம், காக்ஷ்மிரி, மலையாளம், மராத்தி, ஒரியா, பஞ்சாபி, சிங்களம், தமிழ், தெலுங்கு, உருது ஆகிய பதினாங்கு தெற்காசிய மொழிகளின் எழுத்து மற்றும் பேச்சுத் தரவை உள்ளடக்கிய ஒரு மொழி தரவுத்தொகுதிகள் உள்ளன. நுஅடைடந ஒரு மொழி தரவுத்தொகுதி கிட்டத்தட்ட 92,799,000 சொற்களை உள்ளடக்கியது. (இது பெங்காளி, குஜராத்தி, இந்தி, பஞ்சாபி, உருது என்ற மொழிகளின் எழுத்துப்பெயர்க்கப்பட்ட பேச்சுத்தரவின் 2,627,000 சொற்களை உள்ளடக்கியுள்ளது) இணை தரவுத்தொகுதிகள் ஆங்கில உரையின் 2,00,000 சொற்களையும், இந்தி, பெங்காளி, பஞ்சாபி, குஜராத்தி, உருது இவற்றின் மொழிபெயர்ப்புகளையும் கொண்டது. அடையாளப்படுத்தப்பட்ட பகுதி உருது மொழியின் சொல்வகைப்பாட்டு அடையாளப்படுத்தப்பட்ட ஒரு மொழிய மற்றும் இணை தரவுத்தொகுதிகளை உள்ளடக்கியது. இத்துடன் வரையறை பயன்பாட்டின் இயல்பைக்காட்ட இருபது இந்தி தரவுத்தொகுதிகளின் கோப்புகளைக் கொண்டுள்ளது. இந்த விரிதரவானது

CES Compliant SGML- ஐப் பயன்படுத்திக் குறிக்கப்பட்டுள்ளது. மேலும் Unicode-ஐப் பயன்படுத்தி குறியாக்கம் (Encode) செய்யப்பட்டுள்ளது.

3.3.4 EMILLE மற்றும் CIIL ஒரு மொழி எழுத்து தரவுத்தொகுதி

எழுத்து தரவுத்தொகுதி டெல்லியில் உள்ள இந்திய தொழில் நுட்ப நிறுவனம் (CIIT),புவனேசுவரத்தின உள்ள பயன்பாட்டு மொழி அறிவியல் நிறுவனம் (ILAS), அலிகாரில் உள்ள அலிகார் முஸ்லீம் பல்கலைக்கழகம் இவற்றுடன் ஒருங்கிணைந்து இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்டது CIIL தரவுத்தொகுதி ஆகும்.

இந்த தரவுத்தொகுதி தொடக்கத்தில் ISCII உரையாக குறியாக்கம் செய்யப்பட்டுள்ளது. இப்பொழுது CES – Compliant SGML குறியீட்டுடன் Unicode – CIIL ஒரு மொழிய தரவுத்தொகுதி 93,530,000 சொற்களைக் கொண்டுள்ளது.

3.3.5 தமிழுக்கான தரவுத்தொகுதி தயாரித்தல்

தமிழுக்கான தரவுத்தொகுதி தயாரித்தலில் முன்னோடியாக உள்ள நிறுவனம் மைசூரில் உள்ள இந்திய மொழிகளின் மைய நிறுவனம் ஆகும். பல பிற நிறுவனங்களும் தாங்கள் எடுத்துக் கொண்ட இயற்கை மொழியாய்வு திட்டங்களுக்கு ஏற்றவாறு தரவுத்தொகுதிகளைச் சேகரித்துப் பயன்படுத்துகின்றன. மின் வடிவில் உள்ள புத்தகங்களையும் கால இதழ்களையும் நாள் இதழ்களையும் தேவைக்கேற்றவாறு சேகரித்து விரிதரவாகப் பயன்படுத்த இயலும். இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளுக்காக தரவுத்தொகுதி தயாரிக்கும் முயற்சியில் தமிழுக்கு விரிதரவைத் தயாரித்து தந்துள்ளது.

தமிழுக்கான விரிதரவைப் பொறுத்த வரையில் உரைகளானது தினகரன் பத்திரிக்கையின் இணையத்தளத்திலிருந்து எடுக்கப்பட்டுள்ளது. இந்த தரவுத்தொகுதி இந்திய மொழியின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட தொடக்கால விரிதரவையும் உட்படுத்தும். தமிழ் எழுத்து விரிதரவில் எடுக்கப்பட்ட சொற்களின் பட்டியல் கீழே கொடுக்கப்பட்டுள்ளது.

1.சினிமா (cinema) திரைப்படம் மற்றும் அது தொடர்பாக செய்திகள் (கிட்டத்தட்ட 1,050,000 சொற்கள்)

2.செய்திகள் (News) கதைச் செய்திகள் (கிட்டத்தட்ட 8,610,000 சொற்கள்)

3.மற்றவை (Others) இணைய தளத்திலிருந்து எடுக்கப்பட்ட பிற கதைகள் (கிட்டத்தட்ட 730,000 சொற்கள்)

4.அரசியல் (Political) அரசியல் செய்தி அறிக்கை (கிட்டத்தட்ட 5,050,000 சொற்கள்)

5.விளையாட்டுகள் (Sports) விளையாட்டுச் செய்தி அறிக்கை (கிட்டத்தட்ட 1,170,000 சொற்கள்)

பிற நிறுவனங்களின் தமிழ் தரவுத்தொகுதிகள்

இயற்கை மொழி ஆய்வில் குறிப்பாக இயந்திர மொழிபெயர்ப்பில் சூடுப்பட்டுள்ள நிறுனங்கள் தமிழுக்கென்று தரவுத்தொகுதிகள் தயாரித்து அவற்றை சொல் வகைப்பாடுகளுக்கு அடையாளப்படுத்தி பயன்படுத்தி வருகின்றன. குறிப்பாகத் தமிழ்ப்பல்கலைக்கழகம், அண்ணாமலைப் பல்கலைக்கழகம், Aukbc ஆராய்ச்சி நிறுவனம், அமிர்தா தொழில் நுட்ப நிறுவனம் என்பன இயந்திர மொழிபெயர்ப்பு ஆய்வில் சூடுபட்டு அதற்கேற்ற தரவுத்தொகுதிகளை உருவாக்கி பயன்படுத்துகின்றன.

3.4 உரை தரவுத்தொகுதி ஆய்வு

பல மொழிகளில் பெரிய மின் தரவுத்தொகுதிகள் சேகரிக்கப்பட்டபின் அவற்றை ஆய்வு உத்திகளுக்கு உள்ளாக்கும் தேவைவரும். விரிதரவிலிருந்து மொழித்தரவை அணுகுவதற்கும் தேவையான தகவல்களை மீளப்பெறுவதற்கும் மக்கள் ஒழுங்குமுறைகளும் உபாயங்களும் உருவாக்கியுள்ளனர். இவ்வாய்வுக் கருவிகள் பயனுள்ளதாக அமைந்துள்ளன. பலவிதமான ஆய்வு உத்திகள் இருக்கின்றன. எடுத்துக்காட்டாக.

1. புள்ளியல் ஆய்வி (Statistical analyzer)
2. தொடரடைவு ஆய்வி (Concordencer)
3. சொல் சேர்ந்து வருகை ஆய்வி (Lexical collocator)
4. முக்கியச்சொல் காணும் ஆய்வி (Key word finder)
5. அகராதிச்சொல் ஆய்வி (lemmantizer)
6. உருபனியல் ஆய்வியும் உருவாக்கியும் (Marphological analyzer and gernerator)
7. சொல் ஆய்வி (word processer)
8. சொல் வகைப்பாடு அடையாளப்படுத்தி (part of speech tagger)
9. தரவுத்தொகுதி குறிப்புரைப்பான் (corpus annotator)
10. பகுத்துக் குறிப்பான் (Parser)

ஆங்கிலம், பிரஞ்சு, ஜெர்மன் போன்ற மொழிகளுக்கு விரிதரவை ஆய பல மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஆனால் இந்தி, தமிழ் போன்ற இந்திய மொழிகளுக்கு அத்தகைய மென்பொருள்கள் இனிமே தான் உருவாக்கப்பட வேண்டும். இதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

3.4.1 நிகழ்வெண் ஆய்வு

மொழியில் நீண்ட காலமாகவே புள்ளியியலுடனும் கணக்கியலுடனும் தொடர்புடையது. மொழியியல், நடை அளவியல் (stylometrics) போன்ற இயற்கைமொழி விரிதரவிலிருந்து பெறப்படும் வேறுப்பட்ட புள்ளியியல் மற்றும் அளவில் தகவல்களை வேண்டுகிறது. மொழியின் அட்டவணைப் படுத்தப்படும். உரையிலுள்ள பல முக்கியமான மொழி அமைப்பொழுங்குகளை அணுக உதவுவதால் இது மிக முக்கியமானதாகும். இது உள்ளுணர்வால் அறிய இயலாத தகவல்களைக் தருகின்றது.

3.4.2 சொல் வருகை ஆய்வி

சொற்களின் சேர்ந்து வருகை உரைகளில் சொற்களின் பங்களிப்பையும் இடத்தையும் புரிந்துக்கொள்ளப் பெரிதும் உதவும் இவ்வாய்வு சொற்களில் எந்த இணைகளுக்கிடையில் கூடுதல் சேர்ந்து வருகை உறவு இருக்கின்றது என்பதை அறிய இயலும். மேலும் இரண்டு சொற்கள் சேர்ந்து வரும் நிகழ்வுத் தகைமையைக் தற்செயலாக விளைவின் அடிப்படையிலான நிகழ்வு தகைமையுடன் ஒப்பிடும் ஒவ்வொரு சொல் இணைக்கும் ஒரு மதிப்பெண் தரப்படும் கூடுதல் மதிப்பெண் சேர்ந்து வருகையைக்காட்டும். இது அகராதியிலும் தொழில் நுட்ப மொழி பெயர்ப்பிலும் பயன்படுத்த விரிதரவிலிருந்து பல்சொல் அலகுகளைப் பிரித்தெடுக்க உதவும். இயந்திர மொழிபெயர்ப்புக்கும் பொருண்மை வேறுபாட்டிற்கும் இத்தகைய தகவல்களை கைக்கொடுக்கும் இது அர்த்த வேறுபாடுகளை அடையாளம் காண ஒற்றுமையான சொற்களைக் குழும உதவுகிறது.

எ.கா.

மாடிப் - படி Step

வாசற் - படி Step

புத்தகம் படி Read

சொன்னப்படி Manner

ஐந்து லிட்டர் Measure

ஒரே பொருண்மையைக் கொண்ட சொல்லுகளுக்கு இடையே உள்ள பயன்பாட்டு வேறுபாடுகளை வேறுபடுத்த உதவும்.

3.4.3 சூழலில் முக்கியச் சொல்

தரவுத்தொகுதி ஆய்வில் “KWIC” (Key Word in Context (KWIC))’’ என்ற சொல் அடிக்கடி பயன்படுத்தப்படுகிறது. இது குறிப்பிட்ட சொற்களின் நேர்வுகளை அறிய உதவுகின்றது. ஆய்வுக்கு உட்படும் சொல் ஒவ்வொரு வரியின் மத்தியில் கூடுதலான இடை வெளிப்யுடன் தோன்றும் வேறுபட்ட நோக்கங்களுக்கு வேண்டி சூழலின் நீட்சி குறிப்பிடப்பட்டிருக்கும். இது மையத்திலுள்ள சொல்லின் இரண்டு பக்கங்களிலும் இரண்டு மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டும். இதை N-Gram (bi-gram, trigram, tetra-gram set) என்பர் இவ்வமைப்பொழுங்கு ஒருவரின் தேவை அடிப்படையில் மாறும். சொற்கள், தொடர்கள், எச்சத் தொடர்கள் இவற்றின் ஆய்வின் போது கூடுதலான சூழல் அவற்றைப்புரிந்துக்கொள்ளத் தேவைப்படும். மறஇஉ என்பதை ஒரு வரையாக எண்ணுவது நல்லது. சொற்களின் நிகழ்வெண்ணை மையச்சொல்லின் சூழலில் பரிசோதிக்கவேண்டும். எல்லாத் தகவல்களுகம் எப்போதும் தேவைப்படுவதில்லை. ஆனால் நாம் தேவைப்படும் போது தகவலைப் பயன்படுத்துகின்றோம்.

KWIC சொல்லால் பொருந்தலை அணுகியபின் நாம் மொழியியல் வர்ணணையில் பல்வேறுபட்ட நோக்கங்களை உருவாக்கலாம் மற்றும் இந்நோக்கங்களை நிறைவேற்ற வழிமுறைகளை வகுக்கலாம். சூழலில் முக்கியத்துவம், சேர்ந்து வரும் சொற்களின் பங்களிப்பு சூழல்களில் சொற்களின்

உண்மை நடத்தை, நேர்வின் உண்மைச் சூழ்நிலை மற்றும் சூழல் அடிப்படையிலான கட்டுப்பாடுகள் இவற்றைப் புரிந்துகொள்ள KWIC உதவும்.

குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் உரைகளில் சொற்கள் பயன்படுத்தப்படும் அமைப்பொழுங்கைப் பற்றிய தகவல்களை நமக்குத் தரும் மற்றொரு வகையான ஆய்வாகும். இது எங்கு சொல் நிரல் வாக்கியங்களின் பொருண்மைச் சுமையைத் தீர்மானிக்கின்றதோ அங்கும், எங்கு ஒரு உறுப்பின் தனிப்பட்ட பொருண்மைச் சுமை மற்றொரு உறுப்பின் வருகையால் பாதிக்கப்படுகிறதோ அங்கும் முக்கியத்துவம் வாய்ந்ததாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (தொடர்கள் மற்றும் வாக்கியங்களின்) பகுத்துக் குறித்தலின் போது உறுப்புகளின் செயல்பாட்டு நடத்தையை நேரிட தகவலைத் தருகிறது.

இது பயன்பாட்டில் வழக்கமற்ற ஆனால் தற்சுட்டு வடிவுகளுடன் தனிப்பட்ட உறவு கொண்ட வினைவடிவுகளின் எடுத்துக்காட்டாக (Amuse onself, please one self, lend on self remind one self)வகைமுறையைத் தீர்மானிக்கின்றது. இம்மாதிரியான அமைப்பொழுங்கின் அறிவு, மொழி கற்பவர்களுக்கு இடைப்பட்ட நிலையிலிருந்து உயர்ந்த நிலைக்குச் செல்ல குழுமுதலைப் பயன்படுத்தி நாம் தமிழில் முற்றும் வினைகள், பெரும்பாலான எச்ச வினைகள் கூடுதலாகத் தொடரும் என்றும் பெயர்கள் ஒட்டுக்களாலும், பின்னுருபுகளாலும் தொடரப்படும் என்றும் அறிந்து கொள்ளலாம். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல்கள் சொற்களை வினைக் குழுக்களாகவும் பெயர்க் குழுக்களாகவும் ஆய உதவும். குறிப்பிட்ட இடம் சாரந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல் சொற்களின் குறிப்பிட்ட இடம் சார்ந்த சேர்க்கையின் காரணமாக வரும் சொல்மயக்கத்தைத் தீர்க்கின்றது. பொருண்மையின் நுண்மையான வேறுபாடுகள் சூழல்களின் சொற்களின் வருகை முறையுடன் கூடிய உறுப்புகளுக்கிடையே உள்ள அக உறுப்புக்காளல் பெரும்பாலும் தரப்படுகின்றது.

கூட்டுப் பெயர்களுக்கும், கூட்டு வினைகளுக்கும் சொற்களின் குறிப்பிட்ட சேர்க்கையால் உணர்த்தப்படும் பொருண்மைத் தனிப்பட்ட சொற்களின் பொருண்மையிலிருந்து பெற இயலாது.

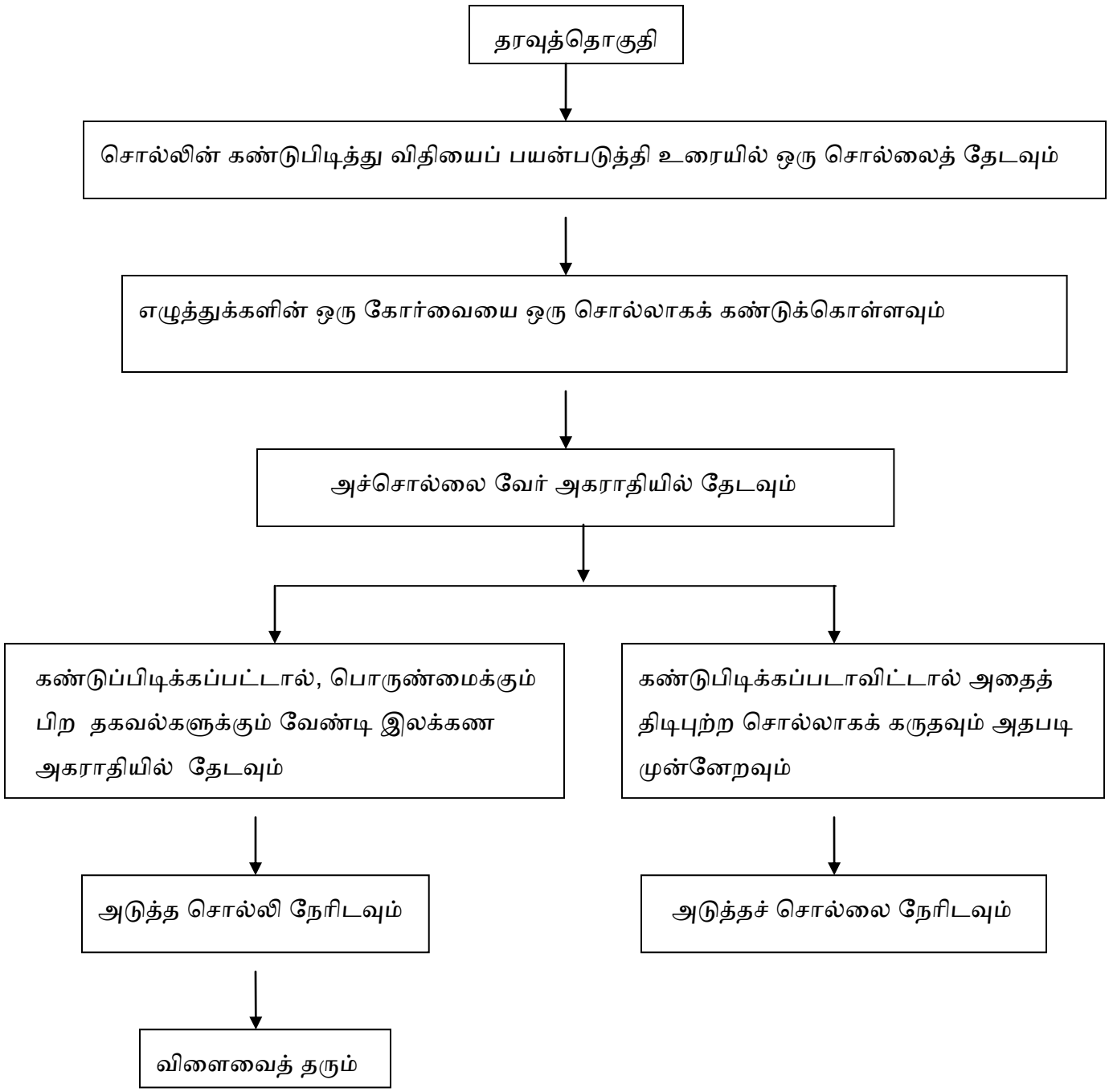
3.4.4 சொல் பகுப்பாய்வு

சொல் பகுப்பாய்வு (Word Processing) விரிதரவில் பயன்படுத்தப்படும் சொற்களைத் தானியக்கமாக ஆய்வதை உள்ளடக்கும். இதன் முக்கிய நோக்கம், ஒரு உரையின் பகுதியிலிருந்து ஒரு சொல்லைக் கண்டுபிடித்து அதை அதன் பயன்பாட்டின் சூழலிருந்து பிரித்து, அதன் உருபொலியனியல் அமைப்பை ஆய்ந்து, அதன் மூலப் பொருண்மையைப் பெற்று, உரையில் அதன் தொடரியல் பங்களிப்பை விளக்குவது அவசியம்.

சொல் பகுப்பாய்விலிருந்து கிடைக்கும் தகவல் சொல் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கும் (Word Sense Disambiguation (WSD))அகராதி உருவாக்குவதற்கும் பகுத்துக் குறிப்பதற்கும் மொழி கற்பத்றகும் மற்றும் இது போன்ற பலற்றிற்கும் பயனுள்ளதாகும். ஆங்கிலம் போன்ற மொழிகளுக்குப் பல சொல் பகுப்பாய்வுகள் உருவாக்கப்பட்டுள்ளன. தமிழ் போன்ற இந்திய மொழிகளுக்கும் இத்ததகைய பகுப்பாய்வு உருவாக்கப்பட்டு பயன்படுத்ததப்பட்டு வருகின்றன.

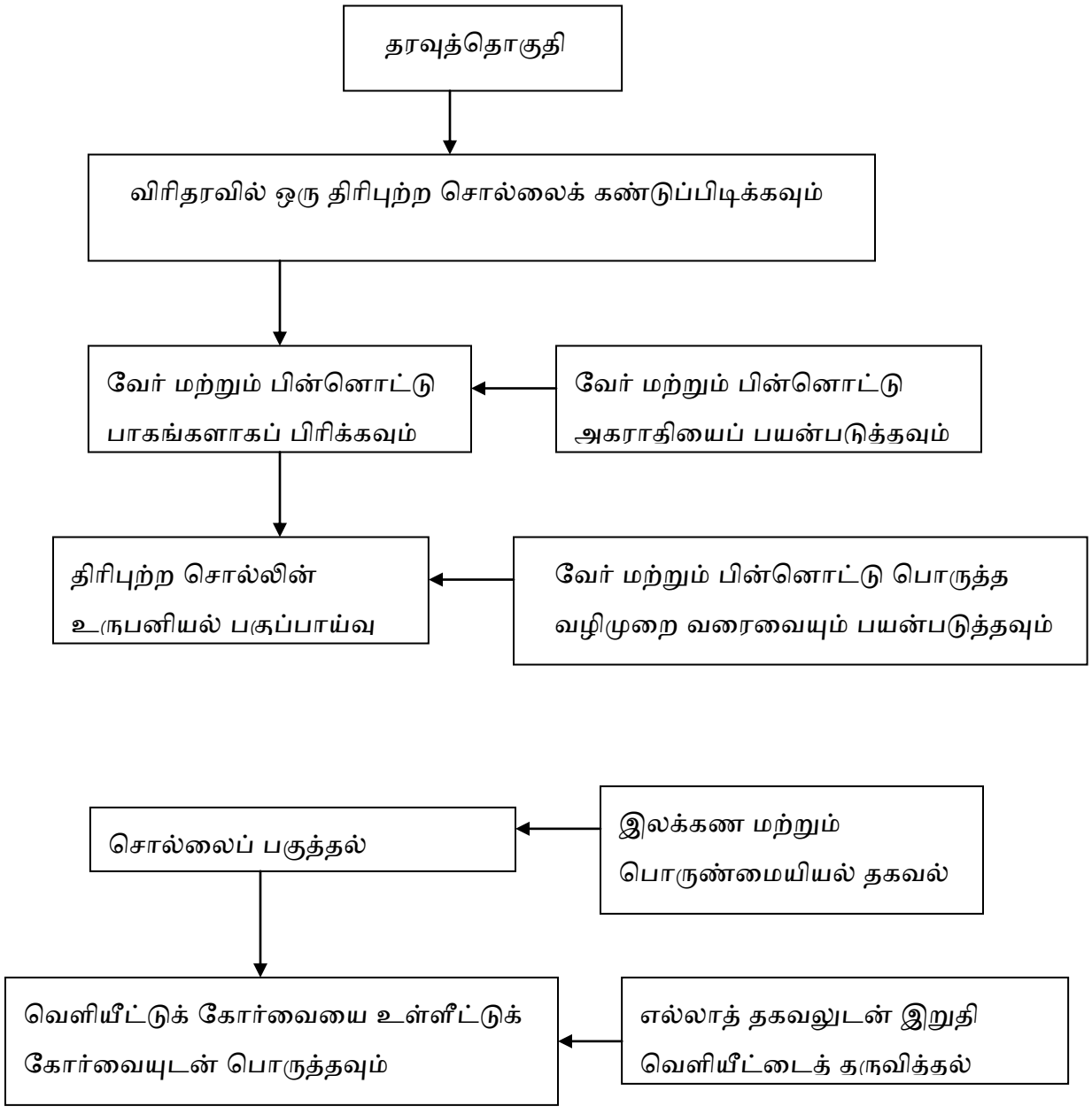
3.4.5 திரிபுறாத சொற்களின் பகுப்பாய்வு

திரிபுறாத சொற்களின் பகுப்பாய்வு (Processing Non-Inflectecd Forms) பின்வரும் படம் சொற்களின் பகுப்பாய்வை விளக்கும்.


தரவுத்தொகுதி
சொல்லின் கண்டுபிடித்து விதியைப் பயன்படுத்தி உரையில் ஒரு சொல்லைத் தேடவும்
எழுத்துக்களின் ஒரு கோர்வையை ஒரு சொல்லாகக் கண்டுக்கொள்ளவும்
அச்சொல்லை வேர் அகராதியில் தேடவும்
கண்டுப்பிடிக்கப்பட்டால், பொருண்மைக்கும் பிற தகவல்களுக்கும் வேண்டி இலக்கண அகராதியில் தேடவும்
கண்டுபிடிக்கப்படாவிட்டால் அதைத் திடிபுற்ற சொல்லாகக் கருதவும் அதபடி முன்னேறவும்
அடுத்த சொல்லி நேரிடவும்
அடுத்தச் சொல்லை நேரிடவும்
விளைவைத் தரும்

3.4.6 திரிபுற்ற சொற்களைப் பகுத்தாய்தல்

பின்வரும்படம் திரிபுற்ற சொற்களின் (Processing Inflecled words) பகுப்பாய்வை விளக்கும்.


தரவுத்தொகுதி
விரிதரவில் ஒரு திரிபுற்ற சொல்லைக் கண்டுப்பிடிக்கவும்
வேர் மற்றும் பின்னொட்டு பாகங்களாகப் பிரிக்கவும்
வேர் மற்றும் பின்னொட்டு அகராதியைப் பயன்படுத்தவும்
திரிபுற்ற சொல்லின் உருபனியல் பகுப்பாய்வு
வேர் மற்றும் பின்னொட்டு பொருத்த வழிமுறை வரைவையும் பயன்படுத்தவும்
சொல்லைப் பகுத்தல்
இலக்கண மற்றும் பொருண்மையியல் தகவல்
வெளியீட்டுக் கோர்வையை உள்ளீட்டுக் கோர்வையுடன் பொருத்தவும்
எல்லாத் தகவலுடன் இறுதி வெளியீட்டைத் தருவித்தல்

3.4.7 இரட்டைச் சொற்களைப் பகுத்தாய்தல்

இரட்டைச் சொற்களின் ஆய்வு (processing dbouble words) உறுப்புகள் அவற்றிற்கு இடையில் ஒரு இடைவெளிப்யால் பிரிக்கப்பட்ட கூட்டுகள் இரட்டுற்ற சொற்கள், வேறுபட்ட சொற்கள் என்பனவற்றை உள்ளடக்கும். இரட்டைசொற்களை விரிதரவில் ஆய்வது மிகவும் சிக்கலான செயல்முறையாகும். பெரும்பாலும் இரட்டைசொற்கள் அகராதியில் தரப்பட்டிருக்கும். அவ்வாறு தரப்படாத சொற்களை காண்பது தான் சிரமமாகும். இதற்கு சேர்ந்து வருகை நிகழ்வெண்கள் இதற்கு பொருந்தும். அடுத்து வருகின்ற பொருண்மை இரட்டைச் சொற்களின் பொருண்மையை உறுதிசெய்ய உயதவுவதால் இரட்டை சொற்களை ஆய்தல் சொல் மட்டத்தில் பொருண்மை மயக்கம் தீர்க்க உதவும்.

3.5 தரவுத்தொகுதி அடையாளப்படுத்தல்

தரவுத்தொகுதி சொல்வகைப்பாட்டிற்கும் இலக்கணத்திற்கும் சொற்பொருளுக்கும் அடையாளப்படுத்தப்படும்.

3.5.1 சொல்வகைப்பாடு அடையாளப்படுத்தல்

சொல் வகைப்பாடு அடையாளப்படுத்தும் (Part Of Speech Tagging) திட்டம் ஒரு சொல்லின் வாக்கியத்தில் அதன் சொல் வகைப்பாட்டுடன் அடையாளப்படுத்தும். இது போன்ற நிலைகளில் செய்யப்படுகிறது. முன்திருத்தல், தானியங்கு அடையாளம் தருகை மற்றும் மனித முயற்சிசார் பின்திருத்தல் என்பவையாகும்.

3.5.2 இலக்கணம் அடையாளப்படுத்தல்

சொல் வகைப்பாடு அடையாளப்படுத்தலில் இரண்டாவது மட்டம் இலக்கணம் அடையாளப்படுத்தல் ஆகும். இது விரிதரவில் உள்ள ஒவ்வொரு சொல்லுக்கும் இலக்கணம் பொருள் தரும் விரிவான குறிப்புரை ஒழுங்கு முறையாகும். இம்முறையில் ஆங்கிலத்திற்கு சில கருவிகள் உருவாக்கப்பட்டுள்ளது. எம்முறையில் செய்திருதாலும் இறுதியில் மனித முயற்சிசார் திருத்தலுக்கும் மதிப்பீட்டிற்கும் திட்டவரைவு செய்யப்பட்டுள்ளது.

3.5.3 சொல் பொருள் அடையாளப் படுத்தல்

சொல் பொருள் அடையாளப்படுத்தல் (word sense tagging) சொல் அடையாளப்படுத்தப்பட்ட மற்றும் இலக்கணம் அடையாளப்படுத்தப்பட்ட உள்ளீட்டு உரையின் சொற்களை ஏற்க வல்லது. தானியங்கு அடையாப் படுத்தலுக்குப்பின் ஒவ்வொரு சொல்லும் சரியான பொருண்மை வகுப்பாக்கத்தை உறுதி செய்யவேண்டி மனித முயற்சி சார் பின் திருத்தல் செய்யப்படுகிறது.

3.6 மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதி

மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பு இன்றைய காலகட்டத்தில் மிக சிறப்பாகவும் பரவலாகவும் பேசப்படும் மற்றும் பயன்படுத்தப்படும் நெறிமுறையாகும். மொழி புரிந்துகொள்ளுதல், பேச்சு புரிந்துக்கொள்ளுதல், உரை மீட்பு மற்றும் புரிந்துக்கொள்ளுதல் உரைகளிலிருந்து தகவல்களை மீட்டல், ஒலிவழி எழுத்து புரிந்துக்கொள்ளுதல் இயந்திரமொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கும். கணிப்பொறி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் கூடி வருகிறது. இருப்பினும் மனித மொழியியல் உட்படும் கலவைத்தன்மைகள் மற்றும் சிக்கல்கள்

காரணமாக மொழி ஆய்வுக்காகத் திட்டமிடப்படும் கணினி வழியமைப்பு முறைகள் வேறுபட்ட மொழித் தரவுகளின் மிகக் கூடுதலான அளவை வேண்டும். இதன் காரணமாக அளவீட்டு மொழியியல் என்று அழைக்கப்படும் தரவுத்தொகுதி மொழியியல் பயன்பாட்டு மொழியியலின் புதிய கிளையாக உருவாக்கியுள்ளது. எல்லா மொழிகளிலும் தரவுத்தொகுதி உருவாக்கப்பட்டு அவற்றை மொழியின் பல நிலைகளிலும் அடையாளப்படுத்தும் முயற்சி நடைப்பெற்று வருகிறது. இத்தகைய தரவுத்தொகுதி மொழிக் கருவிகள் தயாரிப்பதற்கு மிகப்பயனுள்ளதாக அமைகிறது. இந்நெறிமுறை மொழித் தொழில் நுட்பமாக வளர்ந்துள்ளது.

3.6.1 மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதியின் முக்கியத்துவம்

வேறுபட்ட வகையானன தரவுத்தொகுதி உருவாக்கத்தால் மொழித்தொழில் நுட்ப ஆய்வுப் பயன்பாட்டிலும் தரவுத்தொகுதியின் முக்கியத்துவம் பெருகியுள்ளது. மொழி ஆய்வுக்கான கருவிகள் மற்றும் ஒழுங்குமுறைகள் உருவாக்குவதில் மொழி தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி உயர்ந்த விளைவைத் தருகிறது. பொதுவாக நாம் தரவுத்தொகுதியின் பயன்பாட்டை இயந்திரத்தால் கட்டுப்படுத்தப்பட்ட கருவிகளைத் திட்டமிடுவதற்கும் தானியங்கு கருவிகளைப் பரிசோதிப்பதற்குப் பயிற்சி தரவும் சிறப்பான மூலவளமாக காண்கிறோம். விரிதரவிலிருந்து பெறப்பட்ட மொழிப் பண்புக்கூறுகளின் நிகழ்வெண்ணிக்கை மொழி கற்போருக்கு நூல்களைத் திட்டமிடவும் ஒளப்வழி எழுத்துக்களைப் புரிந்துகொள்ளும் (OCR) ஒழுங்குமுறைகளை உருவாக்குவதற்கும் தானியங்கு எழுத்துப்பிழைத் திருத்தங்களை உருவாக்குவதற்கும் பயனுள்ளதாக இருக்கிறது. குறிப்புரை செய்யப்பட்ட மற்றும் குறிப்புரை செய்யப்படாத தரவுத்தொகுதிகள் இயந்திரமொழிப்பெயர்ப்பு ஒழுங்குமுறையை திட்டமிடவும் இயந்திரத்தால்

படிக்கப்பெறும் அகராதிகளைத் (Machine Readable Dictionaries (MRDS)) உருபனியல் ஆய்விகள், சொல் ஆய்விகள், ஆய்விகள், வாக்கியப் பகுத்துக்குறிப்பான்கள் என்பனவற்றை உருவாக்கவும் பயனுள்ளதாக அமைகிறது. கூப்ரிதரவைப் பயன்படுத்தி உருவாக்கப்பட்ட பல கணினி கருவிகளும் ஒழுங்குமுறைகளும் மொழியைப் பயன்படுத்துபவர்களும், ஆய்வாளர்களுக்கும், எழுத்தாளர்களுக்கும, கல்வியாளர்களுக்கும், ஆசிரியர்களுக்கும், மாணவர்களுக்கும், அறிஞர்களுக்கும், அச்சிட்டு வெளிப்யிடுபவர்களுக்கும், மொழி கற்பவர்களுக்கும் மற்றும் பிறருக்கும் மிகுந்த பயனுள்ளதாக அமைகிறது. பொதுவாக தரவுத்தொகுதியின் முக்கியதுவத்தை மொழித் தொழில்நுட்பத்தின் நான்கு விரிந்த களங்களாகப் பகுக்கலாம் : இது அவற்றின் இயல்பு மற்றும் பயன்பாட்டு நோக்கு அடிப்படைடயில் செய்யப்படும்.

- அறிவின் மூலவளமாக தரவுத்தொகுதி
- மொழித் தொழல்நுட்ப கருவிகளையும் ஒழுங்குமுறைகளையும் திட்டமிட மூலவளமாக தரவுத்தொகுதி
- மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி
- மனித இயந்திர இடைமுக (இடையாக) ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி
- அறிவின் மூலவளமாக பெருந்தரவு
- அறிவு மூலவளமாக தரவுத்தொகுதி பின்வரும் செயல்பாடுகளுக்குப் பயன்படுகிறது :
- பன்மொழி நூலகங்களை உருவாக்குதல்
- மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல்
- (அச்சிட்ட மற்றும் மின் வடிவ) ஒரு மொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) இருமொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) பன்மொழி அகராதிகளை உருவாக்குதல்

- (அச்சிட்ட மற்றும் மின் வடிவ) ஒருமொழி சொற்களஞ்சியங்களைப் (பொருட்புல அகராதிகளை) உருவாக்குதல்
- பல்வேறு விதமான (அச்சிட்ட மற்றும் மின் வடிவ) நோக்கீட்டுப் பொருள்கள் உருவாக்குதல்
- இயந்திரத்தால் படிக்கவியலும் அகராதிகளை உருவாக்குதல்
- பன்மொழி சொல் மூலவளங்களை உருவாக்குதல்
- மின் அகராதிகளை உருவாக்குதல்

3.6.2 மொழித் தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்

மொழித் தொழில் நுட்பக்கருவிகள் பின்வருவனவற்றை உள்ளடக்கும் :

- சொல்லாய்வு ஒழுங்முறை
- எழுத்துப்பிழைத்திருத்தம் ஒழுங்குமுறை
- உரை நேர்செய்யும் ஒழுங்குமுறை
- உருபனியல் பகுப்பாய்வுஒழுங்குமுறை
- வாக்கியப் பகுத்துக்குறிப்புஒழுங்குமுறை
- நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை
- சொல் தேடும் இயந்திரம்
- உரை சுருக்கும் ஒழுங்குமுறை
- உரை அடையாளப்படுத்தும் ஒழுங்குமுறை
- தகவல் மீட்கும் ஒழுங்குமுறை
- தொடரடைவு ழுங்குமுறை
- சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை
- சொல் வலை திட்டமிடல்
- பொருண்மை வலை திட்டமிடல்
- சொல் வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை

- ஒரிடம்சார் சொற்களைக் குழுமும் ஒழுங்குமுறை

3.7 மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1.மொழி மூலவளம்பெறும் ஒழுங்குமுறை

2.இயந்திர மொரிபெயர்ப்புஒழுங்குமுறை

3.பன்மொழிபெயர்ப்புஒழுங்குமுறை

4.மொழி கடந்த தகவல் மீட்வுஒழுங்குமுறை

3.8 மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்

2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்

3.உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள்

4.இணையவலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள்

5. கேள்விப் - பதில் ஒழுங்குமுறைகள்

6. கணினியின் உதவியுடன் கட்டளைகள்

7 .கணினியின் தவியுடன் மொழிக் கல்வி

8. உரை உருவாக்கம்

3.9 பேச்சுத் தொழில்நுட்பத்தில் தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1. பேச்சுத் தொழில் நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம்
2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள்
3. தானியங்கு பேச்சு புரிந்துகொள்ளுதல் செயல்பாடு
4 .தானியங்கு பேச்சு உருவாக்கும் ஒழுங்குமுறைகள்
5. தானியங்கு பகுப்பாய்வு ஒழுங்குமுறைகள்
6. தானியங்கு பேசுபவரை அடையானம் காணும் ஒழுங்குமுறைகள்
7. பேச்சுக் குறைபாடுகள் சரிசெய்தல்

3.10 இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி

இயந்திர மொழிபெயர்ப்பு செயற்கை அறிவு நுட்பத்தின் சிக்கலான ஒரு பகுதியாகும். இது மூல மொழியின் வாக்கியங்களிலிருந்து, இலக்குமொழி வாக்கியங்களை உருவாக்குவதற்கான ஒழுங்குமுறையை உருவாக்கவேண்டும். இலக்கு மொழியின் வெளிப்பீடு உயர்ந்த தரமுடையதாக அமையத் தேவையில்லை. ஓரளவுக்குத் தரம் வாய்ந்ததாய் இருந்தால் போதும். இயந்திர மொழிபெயர்ப்பு மொழிகளுக்கிடையே கருத்துப்பரிமாற்றத்திறகும் தகவல் பரிமாற்றத்திற்கும் மொழிகளைக் கடந்த தகவல் மீட்புக்கும் மொழி கற்றலுக்கும் முக்கியத் தொழில் நுட்பமாகும். பல விதமான பயன்பாட்டின் சாத்தியம் சாதாரணமாக இயந்திர மொழிபெயர்ப்பு மின் வணிகம் தகவல் வட்டாரமாக்கம் பல்மொழி ஆவணப்படுத்தல் மற்றும் தகவல் தொழில் நுட்பம் போன்றவற்றில் மிகப் பயனுள்ள தொழில் நுட்பமாக விளங்குகிறது. நாம் இயந்திர மொழிபெயர்ப்புத் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பைப் பற்றிக்

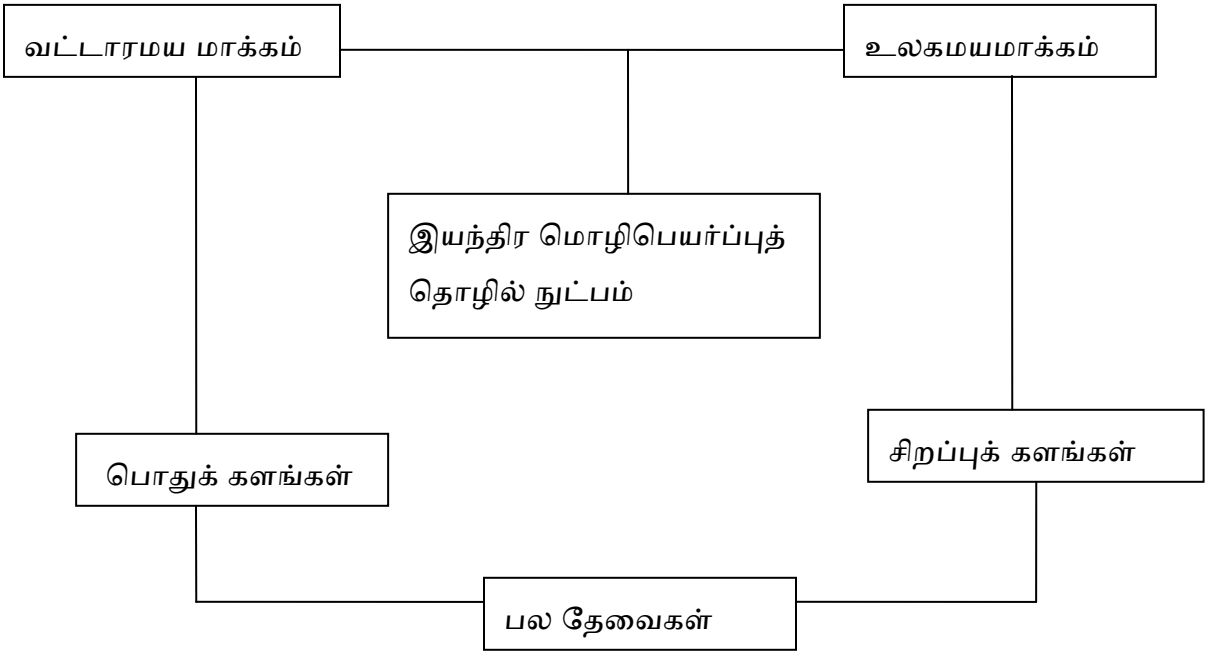
கூறுவதற்கு முன்பின்வருவரும் கேள்விகக்குத் திருப்பதிகரமான விடைதர வேண்டும். மனிதர்கள் மொழி பெயர்ப்பைச் செய்யும்போது ஏன் இயந்திர மொழி பெயர்ப்பை நாம் ஏன் உருவாக்க வேண்டும். இக்கேள்விக்கான பதில் எளிமையானது அல்ல. ஹட்சின் (Hutchin's 1986) பின்வரும் காரணங்களைத் தருகிறார்:

1. பல்வேறுபட்ட அறிவியல் தொழில் நுட்ப ஆவணங்கள் உலகிலுள்ள தொழில் நுட்பவியலாருக்கும் அறிவியலாருக்கும் உடனடியாகத் தரவேண்டுமானால் இயந்திர மொழிபெயர்ப்பு மிகவும் அவசியமாகும்.
2. மனித மொழி பெயர்ப்பாளர்கள் இல்லாத சூழல்களில் இயந்திர மொழிபெயர்ப்பு மிகவும் பயனுள்ளதாக அமையும்.
3. இயந்திர மொழிபெயர்ப்பு மொழகளைக் கடந்த ஆவண மாற்றங்கள் மூலம் அக நிலையான கூட்டுறவை மேம்படுத்த இயலும்.
4. இயந்திர மொழிபெயர்ப்பு பல்வேறுபட்ட அறிவியல், தொழில்நுட்பம், விவசாயம் மற்றும் மருத்துவ தகவல்களை ஏழ்மையான மற்றும் முன்னேறுகின்ற நாடுகளுக்கு விரைவான, எளிதான மற்றம் மலிவான பரிமாற்றம் அடிப்படையில் மொழித் தடைகளை நீக்கி அக நிலையான கூட்டுறவையும் மேம்படுத்த இயலும்.
5. மொழிப் பெயர்ப்பு இராணுவ நோக்கங்களுக்கு தொழில்நுட்ப ஆய்வு மற்றும் வணிக நோக்கங்களுக்கு மிகவும் பயனுள்ளது.

1980-களில் இயந்திர ஹட்சின் இயந்திர மொழிப்பெயர்ப்புக்குச் சாதகமாகக் கூறியபோது இருந்த சூழல் கடந்த ஆண்டுகளில் மிகவும் மாறியுள்ளது. இந்நூற்றாண்டின் தொடக்கத்தில் மக்களின் வாழ்க்கை நடத்தை காரணமாக தாய்மொழி மூலம் தகவலைப் பெறுவது விதியாக மாறிறயுள்ளது. மேலும் படிப்பறிவின் வளர்ச்சி, செய்தி மற்றும் தகவலின் உலகமயமாக்கம், நாடுகளைக் கடந்த பன்மொழிய தரவுத்தொகுதி, இனங்களுக்கிடையே மொழித்

தனித்தன்மையின் அவசியம் மற்றும் பிற காரணிகள் கணினி அறிவியலாரையும் மொழியலாரையும் இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உரவாக்க இணைந்து பணியாற்றும் சூழலுக்குத் தள்ளியுள்ளது.

இக்காரணிகள் பங்களிப்பைப் பின்வரும் வரைபடத்தின் மூலம் சுருக்கமாக கூறலாம்.



இங்கு அனுபவவாத தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு முன்மொழியப்படுகின்றது. இது மொழிபெயர்ப்பில் சூடுபடுத்தபட்டடுள்ள மொழிகளின் உரைகளிலிருந்து உருவாக்கப்பட்ட மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதிகளின் பகுப்பாய்விலிருந்து பெறப்பட்டத் தகவல்களின் மற்றும் எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. இங்கு இயந்திர உதவியால் செய்யப்படும் மனித மொழிபெயர்ப்பிற்கு மொழிபெயர்ப்பு தரவுத்தொகுதி பயன்படுத்தப்பட்ட விவாதிக்கப்படுகின்றது. எல்லைப்படுத்தப்பட்ட மற்றும் பொதுக் களங்களில் இவ்வொழுங்குமுறையின் பயன்பாடு

அழுந்தக்கூறப்படுகிறன்றது. தகவல் பரிமாற்றத்தின் உலக உயர்வுக்கும் வட்டாரவாக்கத்திற்கும் மொழிபெயர்ப்பு ஒழுங்குமுறையின் தேவை விளக்கப்பட்டு எதிர்கால முன்னேற்றத்திற்கு வேண்டிய வழிமுறைகள் முன்வைக்கப்படுகின்றன. மொழி பெயர்ப்பு வரலாற்றின் மற்றும் தோல்வியல் இன்று கற்றுக்கொண்டவைகளின் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு அமைக்கப்பட்ட ஆலோசனை கூறப்பட்டுள்ளது.

3.10.1 நோக்கம்

கடந்த 50 ஆண்டுகளில் இயந்திர மொழிபெயர்ப்பு தொழில் நுட்பங்கள் மற்றும் பயன்பாடு இவற்றின் அடிப்படையில் கூறத்தக்க அளவு வளர்ந்துள்ளது.இன்றைய இணைய தள கால கட்டத்தில் இணைய தகளத்தில் தகவல்கள் பெரியளவில் பல மொழிகளில் உள்ளன. அவற்றை எல்லா மொழி பேசுவோரும் பகிர்ந்து கொள்ளவேண்டுமானால் அவை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு உடனடியாக மொழி மாற்றம் செய்யப்படவேண்டும். இதை இயந்திர மொழிபெயர்ப்பின் விரைவான முன்னேற்றத்திற்ன் அடிப்படையில் தான் நிறைவேற்ற இயலும். இது நடைமுறையில் சாத்தியமில்லாத செயல்பாடாக தற்போது தோன்றினாலும் நாம் இதைச் சாத்தியமாகச் செய்கின்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில் ஆவர்ம் காட்டவேண்டும்.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை ஒரு வடிவிலிருந்து மற்றொருவடிவுக்கு (அதாவது சொல்லிலிருந்து சொல்லையோ கூட்டுச் சொற்களிலிருந்து சுட்டுச் சொற்களையோ தொடரிலிருந்து தொடரையோ வாக்கியத்திலிருந்து வாக்கியத்தையோ) இடப்பெயர்ச்சி செய்யும் அல்லது ஒரு மொழியிலுள்ள உரைகளை மற்றொரு மொழியில் உருப்படுத்தம் செய்யும் எளியமுறை அல்ல. இயந்திர மொழிப்பெயர்ப்பு உருவாக்கப்பட்ட உரை இலக்குமொழியில் இலக்கண அடிப்படையில் சரியான மற்றும் கருத்துரு அடிப்படையில் ஏற்றுக் கொள்ளத்தக்க வகையில் அமைய உறுதி தருவதைக்

குறிப்பிடுகின்றது. இயந்தி மொழிப்பெயர்ப்பு அர்த்தம் மற்றும் பொருளடக்க அடிப்படையில் ஒன்றாக இல்லாவிடினும் குறைந்தது மூலமொழிக்கு அண்மையில் இருக்கவேண்டும். மூல மொழியிலுள்ள தகவல் இலக்கு மொழியில் மாற்றப்படும் போது இழப்பு ஏற்படக்கூடாது. மொழிப்பில் மூலமொழியை விடக்கூடுதல் தகவல் சேர்க்கப்படலாகாது. இயந்திர மொழிபெயர்ப்பு உருபனியல், சொல்லியல், தொடரியல் பொருண்மையியல், பயன்வழியியல், கருத்தாடல், புலனறிவு இவற்றைச் சார்ந்திருக்கவேண்டும். இதன் அர்த்தம் என்னவென்றால் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழிபெயர்ப்பாளருக்குள்ள திறனைப் பெற்றிருக்கவேண்டும் என்பதாகும். இதைச் சாதிப்பதற்கு நாம் காத்திருக்கவேண்டும். இருப்பினும் இருமொழி அல்லது இணை மொழி பெயர்ப்பு தரவுத்தொகுதிகள் இதைச் செய்வதற்குச் சாத்தியமான சூழலைத் தருகிறது. இது நம்மை தரவுத்தொகுதிகள் அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறைக்குக் கொண்டு செல்கிறது. இதன் காரணமாக இயந்திர மொரிப்பெயர்ப்பு ஒழுங்குமுறையின் திறன் கூடுகிறது. குறையான தகவலிலன் மூலமாகத் திருத்தப்படாத வடிவில் பயன்படுத்தப்படுவதன் காரணமாகவும் இயந்திர மொழிப்பெயர்ப்பு முற்றிலும் சரியான மொழிப்பெயர்ப்பை உருவாக்குவதை நோக்கமாகக் கொண்டதல்ல எனக் கூறவியலும். சூயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மொழியில் அடிப்படையிலும் புலனறிவு அடிப்படையிலும் உத்தேச ஒழுங்குமுறையைத்தான் நோக்கமாகக் கொள்கிறது. முயற்சித்தல் மற்றும் பிழை செய்தல் என்ற நீண்ட செயற்பாங்குளின் வழி இலக்கை அடையலாம்.

3.10.2 வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்

1950-களில் இயந்திர மொழி பெயர்ப்பு உலகிலுள்ள பெருமாபாலான எல்லா மொழிகளுக்கும் நேரடியாகவோ மறைமுகமாகவோ பல மூல மொரியிலிருந்து இலக்கு மொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சிகள்

எடுக்கப்பட்டன. இது வரை முழு வெற்றி அடைய இயலவில்லை. அதன் அர்த்தம் என்னவென்றால் மூல மொழிக்கும் இலக்கு மொழிக்கும் மொழி பயன்பாட்டாளர்களால் ஏற்றுக்கொள்ளக்கூடிய மொழிப்பெயர்புகளைத் தானியக்கமாக உருவாக்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் ஒழுங்குமுறை இல்லை. இதன் விளைவாக இயந்திர மொரீபெயர்ப்பு ஆய்வாளர்கள் மரபு சார்ந்த விதி அடிப்படையிலான அணுகுமுறைக்குத் தங்கள் கவனத்தை திருப்பியுள்ளானர்.

3.10.3 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

கடந்த நூற்றாண்டுகளாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கு வேண்டி உருவாக்கப்பட்ட விதிகளின் குழுமங்களால் இயற்கை மொழியின் நெகிழ்வான அமைப்பை உருப்படுத்தம் செய்ய இயலாது. இதன் காரணமாக தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழி பெயர்ப்பு உருவானது. இது பின்வரும் இயந்திர மொழிபெயர்ப்பு வகைகளை உள்ளடக்கும்.

1. எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (example based machine translation)
2. புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (statics based machine translation)

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் விளைவுகளில் ஒன்று இருமொழி உரைகளின் வாக்கியங்களை வரிசைப்படுத்தும் வழிமுறை வரைவுகளின் உருவாக்கமாகும். இவ்வெளிப்ய விளைவு இயந்திர உதவி வழி மனித மொழி பெயர்ப்பின் அடிப்படையில் விக்ஷயங்களில் ஒன்றாக அமைந்தது. ஏனென்றால் இது மொழிபெயர்ப்புக்கு உதவும் பல புதிய கருவிகளுகடகுப் பொருத்தமான அடிப்படையாக அமைந்துள்ளது. மொழிபெயர்ப்பாளர்களால் ஏற்கனவே உருவாக்கப்பட்ட மொழிபெயர்ப்புகளைப்

பயன்படுத்தி முழுமையாகவோ பகுதிகாகவோ அவற்றின் அக அமைப்புகளைக் கண்டுபிடிக்க முயலுகிறது. இப்பகுப்பாவு அடிப்படையிலான அணுகுமுறை மொழி பெயர்ப்பாளர்களுக்கு உதவும் கருவிகளைஉருவாக்க உதவியது. இருமொழி இல்லது இணைமொரி பெயர்ப்பு விரிதரவை இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தும் கருத்து புதியது அல்ல. இயந்திர மொழிபெயர்ப்பின் ஆரம்ப காலத்திலேயே இது முயற்சிக்கப்பட்டது. பொதுவாக இணை தரவுத்தொகுதிகள் ஒரு மொழி தரவுத்தொகுதிகளை விட மொழிகளைப் பற்றிய விரிவான தகவலைக் கொண்டிருக்கிறது. தரவுத்தொகுதி ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறையையும் புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையும் இணைத்து பயன்படுத்துவதால் இது உயர்ந்ததாக இருக்கிறது. விரிததரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மூல மொழியிலும் இலக்கு மொழியிலும் உள்ள தரவுத்தொகுதிகளில் முதன்மையான பகுப்பாயிவிலிருந்து உருவாக்கப்பட்ட வகைகளின் பரப்பெல்லைகளிள் அடித்தளத்தில் அமைந்துள்ளது. பகுப்பாய்வு மொழிபெயர்க்கப்பட்ட விரிதரவில் அடங்கியுள்ள தொடர்கள், மரபுத் தொடர்கள், வாக்கியங்கள், பத்திகள் மற்றும் பிற மொழியியல் பண்புக்கூறுகளின் உருபனியல் பொருண்மையியல் மற்றும் புலனிறவு பொருள்கோள்களை உட்படுத்தும்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடிப்படையிலான நெறிமுறை மொழி பெயர்ப்புக்கு முன் வரையறுக்கப்பட்ட தீர்மானங்கள் இல்லை. ஆனால் மனிதரால் மொழி பெயர்க்கப்பட்ட உரைகளில் மிக சாத்தியமான தீர்மானங்கள் காணப்படும் என்ற நம்பிக்கையின் அடிப்படையில் அமைவதாகும். அதாவது மனித மொழிபெயர்ப்பாளர்கள் திறமையின் பெரும்பகுதி மொழி பெயர்க்கப்பட்ட உரைகளில் காணப்படும் மொழி நிகரன்களில் குறியாக்கம் செய்யப்பட்டுள்ளன என்று கூறலாம். எல்லைக்குட்பட்ட களங்களில் இவ்வணுகுமுறையால் பெறப்பட்ட சமீபகால வெற்றி பொதுவான களங்களிலும் இம்மாதிரியான வெற்றையைப் பெற மொழி சார்ந்த மற்றும்

மொழிக்குப் புறம்பான அறிவுதேவை என்று காட்டுகினறது. எல்லாக் களங்களிலும் இவ்வணுகுமுறையின் வெற்றி குறித்து தீர்மானமாக வருவதுரைக்க இயலாது. இருப்பினும் நன்றாகத் திட்டமிடப்பட்டது மொழிப்பெயர்ப்பு விரிதரவிலிருந்து பெறப்பட்ட தகவலிலிருந்து உருவாக்கப்பட்ட நெறிமுறை எல்லைக்குட்பட்ட மற்றும் பொது வளங்களில் வெற்றயைத் தரும் என்று கூறு இயலும்.

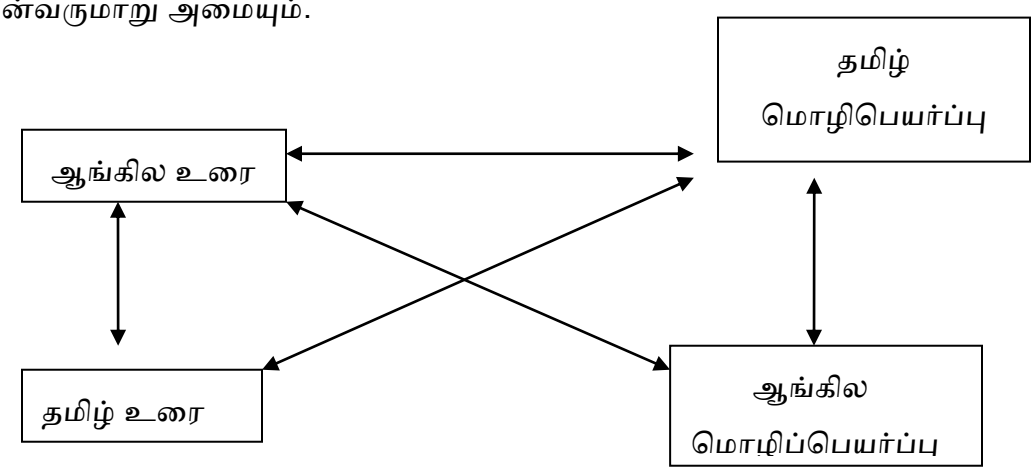
3.10.4 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்

தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு விரிவான தரவுத்தொகுதி பகுப்பாய்வின் மற்றும் ஆய்வின் கூறத்தக்க முன்னேற்றத்தைப் பெற்றுள்ளது. இந்நாள் வரை பல்வேறு வகையிலான மொழிபெயர்ப்பு ஒழுங்குமுறைகள் திட்டமிட உதவும் முக்கியமான உள்ளறிவைத் தரவேண்டி உருவாக்கப்பட்டுள்ளன.. பொதுவாக இம்மொழிபெயர்ப்பு தரவுத்தொகுதி இயற்கையாகப் பெறப்படும் மொழித் தரவுகளின் பெரிய சேகரிப்பை உருப்படுத்தும் செய்கின்றது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழிபெயர்ப்பு தரவுத்தொகுதியின் பயன்பாடு கட்டாயமானதாகும் என்று தெளப்வாக்குகின்றது. ஏனென்றால் அவை மொழி மற்றும் மொழிக்குப் புறம்பான எடுத்துக்காட்டுகளின் மற்றும் தகவல்களின் பல்வேறு விதமான வகைகளை நமக்குத் தருகிறது.

3.10.5 மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கம்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி மூல மொழியின் உண்மையான பண்புக் கூறுகளையும் இலக்கு மொழியிலிருந்து பெறப்பட்ட அவற்றின் மொழி பெயர்ப்புகளையும் கொண்டிருக்கும். இவ்தரவுத்தொகுதி மொழிகளைக் கடந்து சொற்களின் மற்றும் தொடர்களின் பொருண்மையையும் செயல்பாட்டையும்

கொண்டிருக்கும். இதன் காரணமாக ஒரே கட்டுபாட்டிற்குள் இரு வேறுபட்ட மொழிகள் குறிப்பிட்ட பொருண்மைகளின் மெய்யுருவாக்கத்தை ஒப்பிடுவதற்குப் பொருத்தமான அடிப்படையைத் தருகின்றது. மேலும் அவை மொழிகளைக் கடந்த வேறுபட்ட சொற்களைக் கண்டுபிடிப்பதைச் சாத்தியமாக்குகிறது. இவ்வாறு மொழிப்பெயர்ப்பு தரவுத்தொகுதி மொழிகளைக் கடந்த தரவுப் பகுப்பாய்வுக்குத் தருகின்றது. மொழிபெயர்ப்பு தரவுத்தொகுதியின் உருவாக்கம் ஒரு சிக்கலான வேலையாகும். இது தரவுத்தொகுதியின் உருவாக்கம் மற்றும் கவனமான வழிகாட்டலை வேண்டுகிறது. மொழிபெயர்ப்பு தரவுத்தொகுதி ஒப்பீட்டு இணை தரவுத்தொகுதிகளின் முன்னேற்றத்திற்கான வழிகளை ஒன்றிணைக்க வேண்டி தகுதியுடையதாக ஆக்க உருவாக்கப்படும் இரு மொழிகளிலிருந்து உரை மாதிரிகள், உரை வகை, பாடப்பொருள், நோக்கம் மற்றும் கலைச்சொல் இவற்றின் அடிப்படையில் இயன்ற வரைக்கும் பொருத்தப்படும் இரு மொழிகளுக்கிடையிலுள்ள மொழிப்பெயர்ப்பு தரவுத்தொகுதியின் அமைப்பு பின்வருமாறு அமையும்.



இம்மாதிரியான தரவுத்தொகுதிக்கு உரை மாதிரிகள் தெரிவு பின்வருமாறு கொள்கைகளால் பொதுவாக வழி நடத்தப்படும்.

1.எழுதப்பட்ட உரைகளில் பயன்படுத்தப்படும் மொழி மாதிரிகள் மட்டும் தான் மொழி பெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ளன

2.பேச்சு கருத்தாடலிலிருந்து பெறப்படும் உரைகளை உட்படுத்தும் சந்தர்ப்பம் இல்லை. ஏனென்றால் தற்போதைய இயந்திர மொழிபெயர்ப்பு எழுதப்பட்ட உரைகளை மட்டுமே இலக்காகக் கொண்டுள்ளது. உட்படுத்தப்பட்டுள்ள உரைகள் தற்கால மொழியைப் பிரதிபலிப்பது எதிர்பார்க்கப்படும். இருப்பினும், பழங்கால நூல்கள் வரலாற்று உரைகளின் மொழி பெயர்ப்புக்குத் தேவையானதாகும்

3.மொழிப்பெயர்ப்பு தரவுத்தொகுதி எந்தக் குறிப்பிட்ட வட்டார மொழிக்கோ மொழி வகைக்கோ உரிய உரை வகைக்கு எல்லைப்படுத்தப்படவில்லை. அவை மொழி பயன்பாட்டின் எல்லாச் சாத்தியமான களங்கள் மற்றும் துறைகளிலிருந்து பெறப்பட்ட உரை வகைகளின் பரந்த பரப்பெல்லையை உட்படுத்தும்.

4.இரண்டு மொழிகளிலிருந்து எடுக்கப்பட்ட உரைகள் இயன்றவரை ஒப்பிட தக்கவையாக இருக்கும். இவை இனம் (எ.கா.செய்தி), வகை (எ.கா. அரசியல்), பொருளடக்கம் (எ.கா.தேர்தல்) மற்றும் வடிவம் (எ.கா.அறிகை) இவற்றின் அடிப்படையில் பொருத்தம் பார்க்கப்படும். அவைகள் நோக்கம் பயன்படுத்துபவரின் மொழி இவற்றின் அடிப்படையிலும் பொருத்தமுடையதாக இருக்கும்.

5.மொழிபெயர்ப்பு விரிதரவில் உட்படுத்தப்படுள்ள மாதிரிகள் ஒரு இயல்பான தடை நிலையில் (எ.கா. பகுதி, துணைப்பகுதி, அல்லது பத்தி போன்றவை) தொடக்கத்திலிருந்து இறுதிவரை எடுக்கப்பட்ட உரைகளில் விரிவான மற்றும் இயற்கையான பாகங்களைக் கொண்டிருக்கும்.

3.10.6 மொழிபெயர்ப்பு தரவுத் தொகுதியைப் பொருத்தமாக வரிசைப்படுத்தல்

மொழிபெயர்ப்பு தரவுத்தொகுதிகளைப் பொருத்தமான வரிசைப்படுத்துவது என்பது மூலமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு ஆகும். இலக்கு மொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகுடன் பொருந்தவேண்டும் என்பதாகும். இங்கு

மொழிபெயர்ப்பு அலகு என்பது சிறிய தொடர்ச்சிகளான சொற்கள், கூட்டுச்சொற்கள், தொடர்கள் மற்றும் வாக்கியங்கள் என்பனவற்றுடன் பெரிய தொடர்ச்சியான பத்திகள், பகுதிகள் என்பனவற்றையும் உள்ளடக்கும். மொழிபெயர்ப்பு அலகுகளின் தெரிவு மொழியில் ஆய்வுக்குத் தெரிந்தெடுக்கப்பட்டுள்ள பார்வையையும், பயன்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகையையும் பொறுத்து அமையும். மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதி மூலத்திலிருந்து (எ.கா.சட்டம் மற்றும் தொழில் நுட்ப தரவுத்தொகுதிகள்) மாறாத உயர்ந்த நிலையை வேண்டினால் இரு தரவுத்தொகுதிகளின் நெருங்கிய வரிசைப் பொருத்தம் வாக்கியங்கள், சொற்கள் என்ற அளவில் விலகி நிற்கும். தரவுத்தொகுதி தழுவலாக இருந்தால் மூலத்தின் நேரடி மொழிபெயர்ப்பாக அமையாமல் பத்திகள், பகுதிகள் என்ற பெரிய அலகுகளை வரிசைப்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகை அடிப்படையில் நேர்படுத்தப்படும். வரிசையாக வருவதும் மனித மொழிபெயர்ப்பாளர்களின் நம்பகத் தன்மையும் மொழிபெயர்ப்பட்ட தரவுத்தொகுதிகளை வரிசைப்பொருத்தம் செய்ய உதவும். இது கூடுதல் தொழில்நுட்ப தரவுத்தொகுதிக்குப் பகுதி உண்மையாகும். முஹ்றாக, இலக்கிய தரவுத்தொகுதிகள், விரிதரவில் பயன்படுத்தப்பட்டு நிகரன்கள் முன்னரே முறையாக்கம் செய்யப்பட்டிருந்தால் வாக்கிய மட்டத்திற்கு கீழே உள்ள அலகுகளின் வரிசைப் பொருத்தத்திற்கு உட்படும். எந்த நிலையிலும் (பத்தி, வாக்கியம், சொல் போன்ற நிலைகளில்) மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதிகள் இணையான அலகுகள் கொண்ட எளிய சொல் தரவு மையங்களாக் கருதப்படுகிறது. இதன் முக்கிய நோக்கம் இரண்டு மொழிகளுக்கிடையே காணப்படும் அமைப்பு ஒற்றுமைகளைக் காட்டுவதற்கு அல்ல: பயன்வழியில் அடிப்படையில் மூல உரை அலகுகளுக்கு அண்மைப்பட்டதாகக் தோன்றும் இலக்கு உரை அலகுகளைத் தேடுவதாகும். இதைக் செய்வதற்குத் தொடக்கநிலை, இருமொழிய அகராதிகளின் உதவியால் சொற்களின் ஆரம்ப வரிசைப் பொருத்தத்தைச் செய்வதாகும். இம்மாதிரியான

குறை வரிசைப் பொருத்தங்கள் வாக்கிய நிலைகளில் திருப்திகரமான விளைவுகளைத் தரும்.

3.10.7 இந்திய மொழிகளின் இன்றைய நிலை

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பில் முக்கியமான தடைகளில் ஒன்று இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கதபகும். ஒருமொழிய விரிதரவில் போதுமான அளவு எழுதப்பட்ட உரைகள்ள இருந்தாலும் பின்வரும் செயல்படுத்துநிலை சிக்கல்களின் காரணமாக இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளை மறுதிட்டவரைவு செய்வதற்கு ஆர்முள்ள முயற்சிகள் எடுக்கப்பட்டவில்லை.

3.10.8 உருவாக்கம்

இந்திய மொழிகளுக்கிடையே வேறுபட்ட வகையான. மொழிப்பெயர்ப்பு தரவுத்தொகுதி நீண்டநாள் தேவையாகும். சில ஒருமொழிய தரவுத்தொகுதிகள் ஒரேமாதிரியான திட்டவரைவுக் கொள்கைகளையும் உரை வகைகளையும் கொண்டு உருவாக்கப்பட்டுள்ளன. இவற்றை முறையாக மொழிபெயர்ப்பு தரவுத்தொகுதிகளை உருவாக்கப் பயன்படுத்தலாம்.

ஏற்றுக்கொள்ளல்: மனித முயற்சியால் உருவாக்கப்பட்ட மொழிபெயர்ப்பு தரவுத்தொகுதிகள் சில (எ.கா.) (இயவஇடிஇயட bடிடிம வசரளவ) மற்றும் ளுயாவைலய ஆஉயனநஅல போன்றவற்றால் வெளிப்யிடப்பட்ட பல உரைகள்) ஏற்கனவே உருவாக்கப்பட்டுள்ளது என்றாலும் அவை ஏற்றுக்கொள்ளக்கூடிய மின்விடிவில் இல்லை.

மாற்றம்: வேறுபட்ட மூலம், உரைவகை, மாதிரி மற்றும் வடிவமைப்பு கொண்ட பிற தரவுத்தொகுதிகள் இணை மொழிபெயர்ப்பு தரவுத்தொகுதி மாற்ற வேண்டி ஆயப்படவேண்டும்.

சுத்தமாக்கல்: மொழிபெயர்ப்பு தரவுத்தொகுதி எதிர்கால செயற்பாங்கிற்கும் ஏற்றுக்கொள்ளலுக்கு பயன்பாட்டிற்கும் வேண்டி இயந்திரப் பயன்பாட்டு வடிவமைப்பில் வைத்திருக்கமொழி வல்லுனர்களால் முறையாக தலையிடப்படவும் கட்டுப்படுத்தப்படவும் வேண்டும்.

இணைப்பொருத்தமாக வரிசைப்படுத்தல்: மொழிகளுக்குடையில் பொருத்தும் பகுதிகளை அடையாளங்கான வேண்டி மொழிபெயர்ப்பு தரவுத்தொகுதி இணைப்பொருத்தமாக வரிசைப்படத்தப்படவேண்டும்.

3.10.9 மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளில் மொழியின் செயல்பாடுகள்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி தொகுத்தலுக்கும் வரிசைப்படுத்தலுக்கும் பின்னர் மொழியில் பகுப்பாய்வின் வேறுபட்ட நிலைக்கு உள்ளாக்கப்படுவதற்கு முன்னர் பொருத்தமான பல நிலைகளைக் கடக்கும். இது மொழி பெயர்ப்பு நிகரன்கள் முறையாக்கத்தின் வடிவாக்கத்தின் நிலைகளுக்கு அடிப்படையாக அமையும். பொதுவாக மொழியியல் ஆய்வு பின்வருவனவற்றை உள்ளடக்கும் :

- உருபன்களின் வடிவத்தையும் செயல்பாட்டையும் அடையாளம் காண உதவும் சொற்களின் உருபனியல் ஆய்வு
- தொடர்புள்ள தரவுத்தொகுதிகளின் வடிவத்தையும் செயல்பாடுகளையும் அடையாளம் காண உதவும் தொடரியல் ஆய்வு
- தரவுத்தொகுதிகள் பயன்படுத்தப்பட்டுள்ள சொற்களின் புறவடிவங்களுக்கு இடையில் உள்ள இடைமுகத்தை ஆயு உதவும் உருவனியில் - தொடரியல்

ஆய்வு. இதன் துல்லியமான மற்றும் திறமையான ஆய்வு பகுப்பாய்வின் தன்மையையும், விரிவையும் அதிகரிக்கும்.

- சொற்கள், தொடர்கள் போன்ற அலகுகளின் பொருண்மையைக் கண்டுபிடிக்கவும் இதில் வரும் பொருள் மயக்கங்களைக் கண்டுபிடிக்கவும் உதவும் பொருண்மையில்.

3.10.10 மொழிபெயர்ப்பு ஆய்வு

இயந்திர மொழிப்பெயர்ப்பு ஆய்வாளருக்கு இடையிலான முக்கியமான வாக்குவாதம் மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் ஆய்வில் உட்படுத்தம் கலவைத் தன்மையின் நிலையைப் பற்றியாகும். பொதுவான நம்பிக்கை இயற்கையான உரைகளில் வரும் பல எண்ணிக்கையிலான மொழியியல் நடத்தைகள் ஆயப்பட்டு வெளிப்ப்படையாக உருப்படுத்தம் செய்யப்படும் வரை உயர்ந்த தன்மையுள்ள இயந்திர மொழிபெயர்ப்பு எய்த இயலாததாகும் என்பதாகும். சொல் மயக்கம் மற்றும் உறுப்பு பொருத்தம் இவற்றின் சிக்கல்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்டு உட்படுத்தப்பட்ட ஒரு மொழியின் அகராதியிலும் இலக்கணத்திலும் சேகரித்து வைக்கப்பட்டுள்ள ஏராளமான அறிவின் உதவியால் தீர்க்க இயலும். இது மொழிபெயர்ப்பு ஆய்வின் கடுமையான செயற்பாங்கின் பயன்படுத்தலை வேண்டும். சுருக்கமாக இயந்திர மொழிபெயர்ப்பில் மொழி ஆய்வு நுட்பத்திற்குப் பல பயன்பாடுகள் இருக்கின்றன. அவை பின்வருமாறு அமையும்:

- மொழிப்பெயர்ப்பு பகுப்பாய்வு முன்னரே இருக்கக் கூடிய மொழி பெயர்ப்புகளை அமைப்புகளை அமைப்பாக்கம் செய்ய உதவும். இதன் மூலம் அவற்றை புதிய மொழி பெயர்ப்புகளின் உருவாக்கத்தில் மறு உபயோகம் செய்ய இயலும்.

- மொழிப்பெயர்ப்பு பகுப்பாய்வு தொழில் நுட்பம் மொழிப்பெயர்ப்பு தவறுகளின் சில வகைகளைக் கண்டுபிடிக்க வேண்டி தொடக்கநிலை மொழிபெயர்ப்புகளில் பயன்படுத்தப்படுகிறது.
- மொழிப்பெயர்ப்பு பகுப்பாய்வு நுட்பம் இணை மொழிச் சொற்களால் ஏற்படும் குறுக்கீட்டுத் தவறுகளிலிருந்து சுதந்திரமாக மொழிபெயர்ப்பு இருந்தால் அதைப் பரிசோதிக்கப் பயன்படுத்தப்படும்.

3.10.11 இருமொழிய அகராதியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மற்றொரு பயனுள்ள மற்றும் முக்கியப் பகுதி இருமொழி அகராதிகளின் உருவாக்கமாகும். இதன் குறைபாடு இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பின் தடைகளில் ஒன்றாகும். மரபு அகராதியில் இக்குறைபாட்டை நிவர்த்தி செய்ய இயலாது. ஏனென்றால் அவை சொல் துணை வகைப்பாடு, சொல்தெரிவு, சொல்தேர்வுக் கட்டுபாட்டு மற்றும் சொல் அலகுகளின் பயன்பாட்டு களங்கள் இவற்றைப் பற்றிய போதுமான தகவல்களைக் கொண்டிருக்கவில்லை. இருமொழிய அகராதி உருவாக்கத்திற்கு நாம் பல வழிகளில் அடையாளப்படுத்தப்பட்ட மற்றும் பகுத்துக் குறிக்கப்பட்ட தரவுத்தொகுதிகளில் பயன்படுத்தப்பட்டுள்ள பல்வேறு புள்ளியியல் நெறிமுறைகளைப் பயன்படுத்த வேண்டும்.

3.10.12 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு

மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளில் மொழிப்பெயர்ப்பு நிகரன்களைத் தேடல் மூல மொழியில் ஒரு குறிப்பிட்ட பொருண்மையை அல்லது கருத்துருவை வெளிப்ப்படுத்தும் ஒரு குறிப்பிட்ட வடிவிலிருந்து தொடங்கப்பட வேண்டும். மூல மொழியில் தேடல் செயற்பாங்கின் 2-வது பகுதி தொடங்கும். இது இலக்குமொழியில் ஒரே அர்த்தத்தைக் கொண்டிருக்கிற சொல்லின் அடையாளம்

காணலை உட்படுத்தும். இயல்பாகப் பெரும்பாலான மொழிபெயர்ப்பு தரவுத்தொகுதிகள் மொழிபெயர்ப்பு நிகரன்களின் மிகுந்த பரப்பெல்லையை வெளிப்படுத்த இயலும். இவை மாற்று வடிவின் தேர்வை உறுதி செய்யும் காரணிகள், சொற்கள் பயன்பாட்டின் மறு நிகழ்வு அமைப்பொழுங்கின் அடிப்படையில் உறுதி செய்யப்படும்.

3.10.13 கலைச்சொல் தகவல் வங்கியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு தொழில் நுட்ப மற்றும் அறிவியல் கலைச்சொற்களில் பொருத்தமான ஆக்கத்திற்குக் கலைச்சொல் ஆக்கவியலாரின் எண்ணத்தின் அடிப்படையிலான ஆய்வு தேவை.கலைச்சொல் ஆக்கவியலாரின் அடிப்பசைட் செயல்பாடு இலக்கு மொழிக்கு அயல் கருத்துக்களையும் கருத்துருக்களையும் உருப்படுத்தம் செய்ய மிகப் பொருத்தமான அல்லது ஒரளவுக்குப் பொருத்தமான சொற்களைத் தேர்வு செய்வதாகும். இதைச் செய்வதற்குக் கலைச்சொல் ஆக்கவியலார் இலக்கு மொழி பயன்படுத்துபவர்களுக்கிடையில் கலைச்சொற்களின் பொருத்தமான தன்மை, இலக்கணத் தன்மைம, ஏற்றுக்கொள்கை மற்றும் பயன்படு பல்வேறு காரணிகளை மனதில் கொள்ளவேண்டும். முக்கியமான காரணி கலைச்சொற்களின் தரம் ஆகும். இதன்படி மொழியியல் கொள்கைகளின் அடிப்படையில் பொருத்தமான வழிகளைப் பயன்படுத்தி புதிய சொற்கள் உருவாக்கப்டும். மொழிபெயர்ப்பு தரவுத்தொகுதி ஒரு கருத்து, நிகழ்ச்சி, விக்ஷயம் அல்லது கருத்துரு என்பதை உருப்படுத்தம் செய்யப் பலரால் உருவாக்கப்பட்ட பல கலைச்சொற்களின் பெரிய பட்டிலிலிருந்து பொருத்மான ஒன்றைத் தேர்வு செய்யும் முக்கிய பங்களிப்பைச் செய்கின்றது. புதிய சொற்கள் உருவாக்கத்தின் தொடர்ச்சியான முயற்சி தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்பைத் திட்டமிடுவர்களுக்கு ஒன்றுக்குப் பதிலாக ஒன்றைத் தேர்ந்தெடுப்பதில் சிக்கலைத் தரும். புதியச் சொல்லை உருவாக்குவதா அல்லது இலக்கு மொழியில்

பயன்பாட்டால் இயல்பாக்கம் செய்யப்பட்ட மூலமொழிச் சொல்லைப் பயன்படுத்துவதா என்ற விவாதம் எழும். சுப்ல தொழில் நுட்பச் சொற்கள் அவற்றின் உண்மையான மூலத்தைக் கண்டுபிடிக்க இயலாதவாறு இயல்பாக்கம் செய்யப்பட்டிருக்கும். அன்றாடம் வாழ்க்கையில் அவற்றைப் பயன்படுத்தும் உயர்ந்த நிகழ்வெண் அவற்றைச் சொற்றொகையின் பகுதியாக மாற்றும். எனவே மொழிபெயர் பின் போது அவற்றை இடம் பெயர்க்கத் தேவையில்லை. பொதுவாக இலக்கு மொழியில் பெயர்க்கப்பட்ட தரவுத்தொகுதி மூல மொழியிலிருந்து கடன் வாங்கப்பட்டு புதிய கருத்துக்களையும் கருத்துருக்களையும் வெளிப்ப்படுத்தப் பொருத்தமான சொற்களைத் தேர்ந்தெடுக்க நல்ல மூலவளமாகும்.

3.10.14 சொல் தேர்வு கட்டுப்பாடு

மூல மொழியிலிருந்து தெரிந்தெடுக்கப்பட்ட மிகப் பொருத்தமான நிகரன்காளகக் கருதப்பட வேண்டிய இலக்கு மொழிக்கு சொற்களின் தெரிவு மூலமொழியிலும் பட்டறிவுள்ள மொழியியலாரின் திறமையான மற்றும் கவனமான பொரள்கோளை வேண்டும் மற்றொரு கலவைத்தன்மையான செயல்பாடாகும். சூதன் பொருள் என்னவென்றால் மொழியியலார் இலக்கு மொழியில் இருக்கின்ற ஒரே பொருண்மையத் தருகின்ற வடிவங்களின் பெரிய தொகுப்பிலிருந்து கவனத்தில் கொள்ளவும்.

ஆங்கிலம் : that saint ate food

தமிழ் : அந்த முனிவர் போஜனம் செய்தார்

ஆங்கிலம் : he ate food

தமிழ் : அவன் உணவு சாப்பிட்டான்

மேற்சொன்ன எடுத்துக்காட்டுகளில் நாம் தமிழில் எழுவரின் பண்பின் அடிப்படையில் சரியான இணைச் சொல்லைக் தெரிந்தெடுக்க வேண்டியுள்ளது.

மொழியியலாரின் முக்கியமான செயல்பாடு இரு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் கருத்தப்படும் வேறுபட்ட சமூக மொழியியல் காரணிகளுக்குப் பொருத்தமான சொல் அலகுகளைக் காண்பதாகும். இம்மாதிரியான எடுத்துக்காட்டுகள் வெற்றிகரமான மற்றும் அறிவு பூர்வமான மொழிபெயர்ப்பில் சொல் தேர்வு கவனமாகக் கையாளப்பட வேண்டும் என்பதைக் காட்டப்படுகினறது. இச்சிக்கல் மனித மொழிபெயர்ப்பில் கவனமாகக் கையாளப்பட்டாலும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பில் புறக்கணிக்கப்படுகினறன. இச்சிக்கலைத் தீர்ப்பதற்கான நல்ல வழி இயந்திரத்தால் படிக்க இயலும் அகராதியின் தனியான இடத்தில் பொருண்மையில் ஒற்றுமையைக் காட்டும் எல்லா வடிவங்களையும் பட்டியிலிடுவதாகும். இம்மாதிரியான தரவுமையங்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பிரித்தெடுக்க இயலும், பொதுவாக இயந்திரம் படிக்க இயலும் அகராதிகள் பல பாடப்பொருள் களங்களாகப் பிரிக்கப்பட்டிருக்கும். இதனால் ஒரு குறிப்பிட்ட பாடப்பொருளைச் சார்ந்த சொற்களைப் பற்றிய பொருத்தமான தகவலைப் பெற இயலும்.

3.10.15 சொல் மயக்கத்தை நீக்குதல்

இயல்பான சூழலில் மனித கருத்துப் பரிமாற்றம் ஒழுங்கமைப்பு மொழியைக் கருவியாக கொண்டு பேசுபவரிடமிருந்து கேட்பவருக்குச் செய்தியைப் பரிமாற்றம் செய்வதை உள்ளடக்கும். சில நேரங்களில் செய்தியின் பரிமாற்றம் பொருண்மை மயக்கத்திலிருந்து விடுபடாமல் இருக்கும். இப்பொருண்மை மயக்கம் மொழியில் மிகப் பரவலான நடப்பாகும். மயக்கமான சொல்லுடன் தொடர்புடைய அகப்பொருண்மையின் போதாமையால் அல்லது கருத்துப் பரிமாற்றத்தின் ஒரு குறிப்பிட்ட நிகழ்வில் கூற்றின் அமைப்பால் ஏற்படும். இவ்வாறு மயக்கங்கள் சொல் மயக்கம் (lexical ambiguity) என்றும் குறிப்புமரை மயக்கம் (Referential

ambiguity) என்றும் தொடரியல் மயக்கம் (sytacitic mbiguity) என்றும் மூன்றாகப் பகுக்கப்படும்.

they went to bank	-	சொல் மயக்கம்
he love his wife	-	குறிப்புரை மயக்கம்
he waw a girl in part with a binocular	-	தொடரியல் மயக்கம்

இயந்திர மொழிபெயர்ப்பு ஒங்கமைப்பு பேசுபவரின் மன உருப்படுத்தத்தின் புலனுணர்வு அடிப்படையில் உருவாக்கப்படுவதால் இது பேசுபவரால் பயன்படுத்தப்படும் சொற்களுக்கும் வாக்கியங்களுக்கும் எல்லைப்படுத்தப் பட்டிருக்கும். இதை நேர் செய்ய மொழிபெயர்ப்பாளர்கள் மூல அகராதியை ஒரு குறிப்பிட்ட சூழலில் பொருத்தமாக் தோன்றுகின்ற நிகரனை இலக்கு அகராதியால் பொருத்துகிறார்கள். சில நேர்வுகளில் இலக்கு மொழியானது மூலமொழியின் சொல்லை உருப்படுத்தம் செய்ய இயலும் நிகரான சொல் அலகைக் கொண்டிருக்காது. இந்நேர்வுகளில் மொழிபெயர்ப்பாளர்கள்ன சொற்றொடரைப் பயன்படுத்தப் முயல்வர் அல்லது நிகரான விளக்கத்தைத் தர முயல்வர்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பை ஆதரிப்பவர்கள் சொல் மயக்கச் சிக்கலை நீக்க இயற்க்கைமொழி உரைகளில் கூடுதலாக நிகழும் மயக்க வடிவுகளில் பெரும் எண்ணிக்கையை ஆய்ந்தும் வெளிப்ப்படையாக உருப்படுத்தம் செய்தும் உயர்ந்த தரமான மொழிபெயர்ப்பை தரலாம் என்று வாதியிடுகின்றனர். சத்தியம் என்றால் மயக்கமான வடிவுகளும் மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்ட விரிந்த அறிவின் அடிப்படையில் ஆயப்பட்டு இயந்திரத்தால் புரிந்து கொள்ளப்படும் வடிவுகளும் அகராதியில் சேமித்து வைக்கப்பட்ட வேண்டும். ஆனால் பகுத்தறிவாதிகள் இவ்விதம் மொழியியல் தகவல்களைப் பெறும் செயல்பாடு சரியானதும் அல்ல சாத்தியமானதும் அல்ல என்று வாதியிடுகின்றனர். களக்கட்டுப்பாடு

மொழிபெயர்ப்பு (domin – specific translation) மொழிபெயர்ப்புக்கு எப்போதும் ஆழமான பொருண்மையியல் ஆய்வு தேவையில்லை என்று வெளிப்ப்படுத்துகின்றன. குறிப்பிட்ட மயக்கம் உள்ள சொல் அலகுகளின் ஆழமான பொருண்மையியல் ஆய்வுக்குப் பதிலாக நேரடியாக மூல மொழியிலிருந்து இலக்கு மொழிக்குப் பதிலீடு செய்யும் அணுகுமுறையான எளிய ஆய்வைத் தெரிந்தெடுப்பது நல்லது. சில சூழல்களில் மயக்கங்கள் மொழி பெயர்ப்பின் போது அப்படியே கொண்டு செல்லப்படும் என்ற எதிர்ப்பு அடிப்படையில் சில சொல் மயக்கங்களைப் புறக்கணித்தல் சாத்தியமானதும் தேவையானதுமாகும். இருப்பினும் சொல் மயக்கங்களின் ஆய்வு, இலக்கு மொழியில் மயக்கமற்ற உருப்படுத்தங்களை உருவாக்கும் என்ற நிலையில் பொதுவான மொழிபெயர்ப்பு நேர்வுகளில் மயக்கத்தைப் புறக்கணிக்கக் கூடாது.

3.10.16 இலக்கணப் பொருத்தம்

மொழியில் மொழிபெயர்ப்பில் இலக்கணப் பொருத்தம் என்பது மூல மொழியில் ஒரு சொல் இலக்கு மொழியில் உள்ள சொல்லுடன் அர்த்தம் உள் மொழிபெயர்ப்பைப் பெற வேண்டி பொருத்தப்படும் மொழிபெயர்ப்பில் பொருத்தத்திற்கு வேறுபட்ட திட்டங்கள் இருக்கின்றன. அவை சொல் நிலைப் பொருத்தம், உருபன் நிலைப் பொருத்தம், இலக்கணப் பொருத்தம், தொடர்நிலைப் பொருத்தம், வாக்கியநிலைப் பொருத்தம் என்பனவாகும். மிகப் பொதுவானது மொழிபெயர்ப்புக்கு எடுத்துக்கொள்ளப்ட்ட இரு மொழிகளுக்கிடையில் வினை பங்களிப்பைப் பொருத்துவதாகும. இலக்கண பொருத்தம் வகைப்பாட்டியியல் அடிப்படையில் வேறுபட்ட இரண்டு மொழிகளுக்கிடையில் செய்யப்படும் இயந்திர மொழி பெயர்ப்பில் முக்கியத்துவம் உள்ளதாகும். தற்போதைய சூழலில் நாம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்ப்பைப் பற்றி பேசும் பொது இது முக்கியத்துவம் வாய்ந்தது: என்னவென்றால் ஆங்கிலத்திற்கு SVO அமைப்பு உண்டு He is food தமிழுக்கு SOV அமைப்பு இருக்கிறது (அவன் சோறு

சாப்பிட்டான்). எனவே இலக்கணப் பொருத்தமும் சொல் வரிசைமுறை மாற்றமும் இலக்கு மொழியில் வெளிப்யீடுகளை வெளிப்டுவதற்கு முக்கியத்துவம் வாய்ந்ததாகும்.

இயல் 4

தமிழ்த் தரவுத்தொகுதியைச் சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தல்

4.1 அறிமுகம்

தரவுத்தொகுதிகளைச் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்துதல் முக்கியமான செயல்பாடாகும். சொல்வகைப்பாடு அடையாளப் படுத்தப்படாத தரவுத்தொகுதிகள் இயற்கை மொழி ஆய்வுக்கு பயன்படுத்த முடியாது. குறிப்பாக இயந்திர மொழிப்பெயர்ப்புக்குச் சொல்வகைப்பாடு அடையாளப்படுத்துதல் மிக முக்கியமானதாகும்.

தமிழ்த் தரவுத் தொகுதிகளை சொல்வகைப்பட்டிறிகு அடையாளப் படுத்தல் பற்றி குறிப்பாக இந்திய மொழியிலிருந்து இந்திய மொழிகளின் மொழிபெயர்ப்பு ஒழுங்குமுறையில் தமிழ்த் தரவுத் தொகுதி எவ்வாறு அடையாளப்படுத்தப்படுகின்றன எச்சொல்வகை அடையாளங்கள் பயன் படுத்தப்படுகின்றன என்பன குறித்து இங்கு விரிவாக விளக்கப்படும்.

4.2 தரவுத்தொகுதியை அடையாளப்படுத்தல்

தரவுத்தொகுதி பல மொழியியல் தகவல்களுக்காக அடையாளப் படுத்தப்படுகின்றன. சொல்வகைப்பாடு அடையாளப்படுத்தல் (Parts of Speech

Annotation), மீக்கூறு அடையாளப்படுத்தல் (Prosodic Annotation), முன்வரு கிளவி அடையாளப்படுத்தல் (Anaphoric Annotation), பொருண்மை அடையாளப்படுத்தல் (Semantic Annotation), கருத்தாடல் அடையாளப்படுத்தல் (Discourse Annotation) என்பன முக்கியமான தரவுத்தொகுதி அடையாளப்படுத்தல்களாகும் (Annotation of corpus).

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆய்வுக்கு மிகப் பயனுள்ளதாக அமையும். இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி சொற்கள் சொல்வகுப்புகளாக அடையாளப்படுத்தப்பட்டிருக்கும் தரவின் மிகப் பொதுவான வடிவாகும். Brown Corpus, Lorcorpus, BNC என்பவை இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் ஆகும். LLC (Lond Lund Corpus of Spoken English) மீக்கூறுகளுக்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியாகும். Suffanne corpus தொடரியலுக்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆகும்.

4.2.1 சொல்வகைப்பாடு அடையாளப்படுத்தல்

சொல்வகைப்பாடு அடையாளப்படுத்தலின் (Parts of Speech Annotation) நோக்கம் ஒரு பனுவலில் உள்ள ஒவ்வொரு சொல் அலகுக்கும் அதன் வகைப்பாட்டைக் காட்டும் ஒரு கூறிலே தருவதாகும். இது விரிதரவிலிருந்து தரவை மீட்டெடுப்பதில் தனித் தன்மையைக் கூட்டுகிறது. இது தொடரியல் பகுத்து அடையாளப்படுத்தலுக்கும், பொருண்மைக் கள அடையாளப்படுத்தலுக்கும் உதவுகிறது. இது ஒற்றுமையான எழுத்துக்களை வேறுபடுத்துவதற்கு நம்மை அனுமதிக்கின்றது. சொல்வகை அடையாளப்படுத்தல் பின்வரும் தன்மைகள் கொண்டதாகும்:

- ஒவ்வொரு சொல்லுக்கும் ஒரு சொல்வகைப்பாட்டை அல்லது பிற சொல்வகை அடையாளத்தைத் தருவதாகும்.
- அடையாளப் படுத்துகைப் பொதுவாக நிறுத்தற்குறிகளுக்கும் பயன்படுத்தப்படும்.
- கணினிமொழியில் அடையாளப் படுத்துதல் [tokenization] போல் தான் இயற்கை மொழியிலும் சொல்லை அடையாளப்படுத்தும் செயல்பாடு அமையும்.
- சொல்வகை அடையாளப்படுத்தல் பேச்சுப் புரிந்து கொள்ளுதல் [Speech Recongition] இயற்கை மொழி பகுத்துக் குறித்தல் [Natural Language Passing]மற்றும் தகவல் மீட்டுப் பெறல் [Information retrival] என்பனவற்றில் முக்கியமான பங்கு வகிக்கின்றது.
- சொல்வகை அடையாளப்படுத்தலின் உள்ளீடு [Input] வழிமுறை வரைவு [algorithm] சொற்களின் கோர்வைகளும் குறிப்பிட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமமும் [Part of Speech Tagsets] ஆகும்.
- விடுவரல் [Output] ஒவ்வொரு சொல்லுக்கும் ஒவ்வொரு நல்ல அடையாளத்தைத் தருவதுதான்.

  எ.கா:

  அவன் <மா.பெ> என் <வ.மா.பெ> சொல் <வி.பெ> படி <வி.பெ.பி.உ> கேட்டான்.

POS அடையாளப்படுத்தியின் முக்கியமான வேலை சூழலுக்குத் தக்கவாறு சரியான அடையாளத்தைத் தேர்ந்தெடுத்து மயக்கத்தைத் தீர்ப்பதுதான். பெரும்பாலான தமிழ்ச் சொற்கள் எல்லாம் மயக்கம் இல்லாதவை. இருப்பினும் பொதுவாகப் பயன்படுத்தப்படும் சொற்கள் மயக்கம் உள்ளவை.

4.2.2 முன்வருகிளவி அடையாளப்படுத்தல்

முன்வருகிளவி அடையாளப்படுத்தலில் (Annaphic Annotation) எல்லா மாற்றுப் பெயர்களும் பெயர்த்தொடர்களும் அகப்பற்றாக (Cohesion) என்பதன் விரிவான சட்டகத்திற்குள் ஒரே குறிபடு பொருளைக் குறிப்பதற்காக அடையாளப்படுத்தப்படுகிறது. இங்கு முன்வருகிளவியும் வேறுபட்ட வகைகள் வகைப்படுத்தப்பட்டு வரிசைப்படுத்தப்படுகின்றன. இந்த முன்வருகிளவி அடையாளப்படுத்தப்படுத்தல் திட்டம் மாற்றுப்பெயர் தீர்வு (Anaphora resolution) போன்ற ஆய்வு மற்றும் பரிசோதனைச் செய்யு[ம் இயக்க நெறிமுறைகளுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா.

ரவி தன் தங்கையைப் பார்க்கப்போகும் ராமுவைச் சந்தித்தான்

இது பனுவல் புரிந்து கொள்கைக்கும், இயந்திர மொழிபெயர்ப்புக்கும் முக்கியமானதாகும்.

4.2.3 மீக்கூறு அடையாளப்படுத்தல்

மீக்கூறு அடையாளப்படுத்தலின் (Prosodic Annotation) இசையோட்ட ஒழுங்மைப்புகள், அழுத்தம் மற்றும் விட்டிசை என்பன பேச்சில் அடையாளப் படுத்தப்படும் இது மிகச் சிரமமான அடையாளப்படுத்தலே ஆகும். ஏனெனில் மீக்கூறு பிறமொழி நிலைகளை விட இயல்பில் தன் உணர்ச்சி அடிப்படையில் அமையும். இது பயிற்சி பெற்ற காதால் கவனமாகக் கேட்கப்படுத்தலைக் கேட்கும் Lancaster/IBM Spoken English Corpus மீக்கூறுக்காக அடையாளப்படுத்தப்பட்ட விரிதரவாகும். இதில் அழுத்தப்பட்ட அசைகள் சுதந்திரமான இசைமை இயக்ககத்துடன் மற்றும் இயக்கம் இல்லாமல் வேறுபட்ட குறியீடுகளால்

அடையாளப்படுத்தப் பட்டுள்ளன. இசைமையைச் சூழலிலிருந்து ஊகிக்க இயலும். அழுத்தம் இல்லாத அசைகள் அடையாளப்படுத்தப்படாமல் விடப்பட்டுள்ளன.

4.2.4 பொருண்மை அடையாளப்படுத்தல்

பொருண்மை அடையாளப்படுத்தலில் (Semantic Annotation) பனுவலில் உள்ள சொற்களின் பொருண்மைப் பண்புக் கூறுகளோ (குறிப்பாக சொல் அர்த்தத்தின் அடையாளப்படுத்துகை) அல்லது சொற்களுக்கு இடையேயுள்ள பொருண்மை உறவுகளோ (குறிப்பிட்ட செயலின் செயல்கள், பாதிப்பவர்கள் போன்றவை) அடையாளப்படுத்தப்படுகின்றது.

எ.கா.

அவள் பழம் சாப்பிட்டாள்

அவள் - (படர்க்கை பெண்பால், ஒருமை, மாற்றுப்பெயர்,பதிலீடுபெயர்

பழம் - (தாவரத்தின் பூவிலிருந்து விளையும் விதையுள்ள இனிப்பான பொருள்

சாப்பிட்டாள் - (பழத்தை உட்கொள்ளுதல்)

எந்தப் பொருண்மைப் பண்புக் கூறுகள் அடையாளப்படுத்தப்பட வேண்டும் என்பதில் ஒருங்கிணைந்த கருத்து இல்லை. சிலர் Roget's Thesaurus –இல் உள்ள பொதுமையான பொருண்மை வகைப்பாடுகளை முன் வைத்தனர். இம்மாதிரியான அடையாளப்படுத்தும் திட்டம் சொற்களின் மூடப்பட்ட வகுப்புகளையும் திறந்த வகுப்புகளையும் நடைமுறைப்படுத்த அடையாளப்படுத்தப் படுகின்றது.

4.2.5 கருத்தாடல் அடையாளப்படுத்தல்

கருத்தாடல் அடையாளப்படுத்தலில் (Discourse Annotation) ஒரு தரவுத்தொகுதி பனுவல் மற்றும் கருத்தாடல் நிலைகளில் அடையாளப்படுத்தப்படுகின்றது. அடையாளப்படுத்துகின்ற வகை கருத்தாடல் ஆய்வில் முக்கியப் பங்களிப்பைச் செய்தாலும் பரவலாகப் பயன்படுத்தப்படுவதில்லை. ஏனெனில் மொழி வகைப்பாடுகள் சூழல் சார்ந்தவை மற்றும் அவற்றைப் பனுவல்கிளல் அடையாளம் காணுதல் பிற மொழியியல் நடப்புகளைக் காட்டிலும் விவாதத்திற்கு கூடுதல் மூல காரணமாக அமையும். Londond – Lund – Spoken Corpus–இல் வயதுக்கு வந்தவர்களின் பேச்சு போக்கைக் கவனிக்க 16 கருத்தாடல்கள் பயன்படுகின்றது.

4.2.6 பகுத்தல்

பகுத்தல் என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்கப் பகுப்பாய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும். வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுத்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளைக் கொண்டு வருகிறது. வாக்கியநிலைப் பகுத்தல் சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree bank) என்று அழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி

அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா:

'Pearl sat on a chair' என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – sat VVI] [PP on II]
[NP a – ATI] [chair – NNI]
[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்துக் குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.
2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பதுத்துக் குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகின்றது. பகுத்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்தலின் குறைப்பாடு பகுத்தல், தரவுத்தொகுதி பிரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள் சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

4.2.7 தலைச்சொல்லாக்கம்

தலைச்சொல்லாக்கம் (Lemmatization) செயல்பாடு விரிதரவில் பயன்படுத்தப்பட்டுள்ள சொற்களின் மூல வகைப்பாட்டைக் கண்டுபிடிப்பதும் சொற்களை அவற்றுடன் தொடர்புடைய சொல்லன்களாகக் (Lexeme) குறைப்பதும் ஆகும். எடுத்துக்காட்டாக *படித்தான்*, *படிக்கிறான்*, *படிப்பான்*, *படிப்பாள்* என்ற சொல் வடிவுகளிலிருந்து *படி* எனக் குறைத்தல். இது சொல்லின் எல்லா மாற்று வடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச் சொல்லின் மாற்றுவடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச்சொல்லின் மாற்று

வடிவங்களைப் பிரித்தெடுக்கவும், பரிசோதிக்கவும் ஆய்வாளர்களை அனுமதிக்கின்றது. இது மொழி கற்றலுக்குப் பயன் உள்ளதாக அமைகின்றது. இங்கு கற்பவர்கள் ஒரு தலைப்புச் சொல்லின் மொத்த சாத்தியமான எண்ணிக்கையில் பயிற்சி பெறுகின்றார்கள். இது எந்தச் சொற்கள் திரிபுறுகின்றன என்பதை அறிவதற்குப் பயன் உள்ளதாக இருக்கின்றது. எடுத்துக்காட்டாக Brown Corpus–இன் ஒரு பகுதி சொல் மற்றும் இலக்கணத் தகவலுடன் கூடிய சொற்களின் தலைப்பாக்கம் செய்யப்பட்ட வடிவுகளைக் கொண்டிருக்கின்றது.

4.3 சொல்வகை அடையாளப்படுத்தலின் முக்கியத்துவம்

மொழி ஆய்வில் சொல் வகைப் பாட்டின் முக்கியத்துவம் என்னவென்றால் இது சொல்லைப் பற்றியும் அவற்றை அடுத்து வரும் சொற்கள் பற்றியும் செய்திகளை ஏராளமாகத் தருகின்றது. இவை முக்கிய வகைப்பாடுகளுக்கு உண்மையாகும். வினை-பெயர் மட்டுமின்றி பல நுட்பமான வேறுபாடுகளுக்கும் உண்மையாகும். எடுத்துக்காட்டாக இந்த அடையாளக் குழுமங்கள் உடமை மற்றும் பெயர்களுக்கும் (*என்னுடைய, உன்னுடைய, அவனுடைய, அவளுடைய, அவருடைய, அதனுடைய*) மூவிடப் பெயர்களுக்கும் (*நான், நீ, அவன், அவள், அது*) வேறுபாட்டைக் காட்டுகின்றன.

ஒரு சொல் உடமை மாற்றுப்பெயரா? அல்லது மூவிடப்பெயரா? என்பதை அறிந்துகொண்டு எந்தச் சொற்கள் அதன் பக்கத்தில் வர இயலும் என்பதைக் கூற இயலும். இதைப் பேச்சுத் தெரிந்துகொள்ளுதல் (Speech Recognition) என்பதற்கான மொழி மாதிரியில் (Language Model) பயன்படுத்தலாம். சொல்வகைப்பாடு தகவல் மீட்டலுக்கான (Information retrieval) வேர்ச்சொல்லைக் கண்டுபிடித்தலுக்குப் பயன்படுகின்றது.

ஏனென்றால் சொல்வகைப்பாட்டை (Parts of Speech) அறிவது அவை எந்த உருபனியல் ஒட்டுகளையும் ஏற்கும் என்பதை அறியலாம். அவை பனுவலிலிருந்து பெயர்களையோ முக்கிய சொற்களையோ அறிய உதவுவதால் தகவல் மீட்டல் பயன்பாட்டிலும் பயன்படுத்தலாம். தானியங்கு சொல்வகைப்பாட்டு அடையாளப் படுத்திகள் சொல்பொருள் மயக்கம் நீக்கி வழியமைப்பு முறை (word sense disambiguation algorithms) சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (Part-of-speech-taggers=pos taggers), வகுப்பு அடையாளப்படுத்திகள் (class based N-grams) போன்ற முன்னேறிய ASR மொழி மாதிரிகளில் பயன்படுகின்றது.

ஆங்கிலத்தில் புகழ்பெற்ற POS tagger Brown Corpus-இல் பயன்படுத்தப்பட்ட 87 அடையாளக் குழுமங்களிலிருந்து (tagsets) உருவாக்கப்பட்டவை. மூன்று மிகப் பொதுவாகப் பயன்படுத்தப் படுகின்றன.

1. 45--Tag Penn Tree bank tagset.
2. The medium Size 61 tag 65 tagsets used by Lancaster VCREL, Projects CLAWS (Constituent Liklihood Audomatic Word-Tagging System) Tagger to tag British National Corpus (BNC)
3. The Longer 146-tag C7 tagset

Ex.

Penn Tree bank Version of Brown Corpus

ASCII Flat file

Tag represented after each word following a slash

The/DT grand /JJ/ Jury/NN commented/VBD on/IN a/DT number/NN of/IN others/JJ Topics/NNS%

DT – Deter miners

JJ – Adjective

NN – Noun,sing,or mass

NBD – Verb, Past tense

IN – Preposition

NNS – Noun plural

4.4 சொல் வகைப்பாட்டை அடையாளப்படுத்தலின் வகைகள்

மூல சொல்வகை அடையாளப்படுத்துவதின் வழிமுறை வரைவு (Algorithm) இரு வகுப்புகளில் அடங்கும். மூன்று அடையாளப்படுத்திகள் (taggers) பயன்பாட்டடில் உள்ளன. அவையாவன:

1. விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Rules based taggers)
2. புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochatic Taggers)
3. மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள் (Transformation based taggers)

4.4.1 விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் மயக்கம் தருகின்ற சொல் பெயரா?, வினையா? என்று குறிப்பிடுகையால் எழுதிய பெரிய தரவு மையங்களை [data base] உள்ளடக்கியது. ஆரம்பக் கட்டத்தில் தானாகவே சொல்வகைப்பாட்டை அடையாளப்படுத்துகின்றது. அவ்வாறு அடையாளப்படுத்த பயன்படுத்தப்பட்ட வழிமுறை வரைவுகள் (Algorithm) இரண்டு நிலை அமைப்பால் ஆனதாகும்.

1. அகராதி ஒவ்வொரு சொல்லுக்கும் சாத்தியமான சொல்வகைப்பாட்டைத் தரும்.
2. இரண்டாம் நிலை கையால் எழுதப்பட்ட மயக்கம் தரும் சொல்வகைப்பாட்டைத் தருமாறு பயன்படுத்தும்.

எடுத்துக்காட்டாக ENG Two Level Tagger இரண்டாம் நிலை அமைப்பால் ஆனதாகும். ENG TWO Level அகராதி இருநிலை உருபனியல் [Two level Morphology] அடிப்படையில் அமைந்தது.

4.4.2 புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochostic Part of Speech Tagging)

சாத்தியங்கள் அடையாளப்படுத்துவதில் பயன்படுத்துவது மிகப் பழமையானது. பல சாத்திய அடிப்படையிலான அடையாளப்படுத்திகள் வழக்கில் உள்ளன. எ.கா. Hidden Markove Model or HMM Tagger. இந்தப் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தியில் பயன்படுத்தப்படும் உள்ளுணர்வு 'ஒரு சொல்லின் கூடுதலான சாத்தியமான அடையாளத்தை எடு' [Pick up the most likely tag for the Word] என்பதாகும். தரப்பட்ட வாக்கியத்திற்கு அல்லது சொல் கோர்வைக்கு HMM அடையாளப்படுத்தி பின்வரும் வாய்ப்பாட்டைப் (formula) பெரிதுபடுத்தும் அடையாளக் கோர்வையைத் தேர்ந்தெடுக்கின்றது.

P(wird/Tag) *P(Tag/Previous n tag)

HMM ஒரு சொல்லைக் காட்டிலும் ஒரு முழு வாக்கியத்திற்கு அடையாளக் கோர்வையைத் தருகின்றது. இவ்வகையான Bigram – HMM Tagger Wi என்ற சொல்லுக்கு ti என்ற அடையாளத்தைத் தேர்ந்தெடுக்கும். அது முந்தைய

அடையாளம் ti–ஐ மற்றும் தற்போதைய சொல் wi என்று தரப்படுகையில் கூடுதலான சாத்தியமான அடையாளத்தைத் தேர்ந்தெடுக்கும்.

ti = arg max P(tj/ti-l,wi)

j

ஒரு அடையாளப் படுத்தலுக்கு அடிப்படையான HMM சமன்பாடு

ti argmax (tj,Pi-l) P(witj)

j

எ.கா:

(1) Secretariat,NNP is /VB2 expected/VBN to/To race VB Tomorrow/NN

(2) people/NNs continue NBP to/To inquire/VB The DT Reason/NN for/in The/CT race/NN For/in outer/JJ Space/NN.

முதல் எடுத்துக்காட்டில் race ஒரு வினை (VB) இரண்டாவது எடுத்துக்காட்டில் race ஒரு பெயர் (NN) Bigram HMM அடையாளப்படுத்தி அடையாளச் சிக்கலை அண்மையிலுள்ள சொற்களையும், அடையாளங்களையும் பார்த்துத் தீர்மானிக்கலாம் என்று ஊகம் செய்கின்றது.

To/To race/?????

Tge/DT race/????

எவ்வாறு race என்ற நமது எடுத்துக்காட்டு பயன்படுகின்றது என்று பார்ப்போம். நாம் NN -க்கும் VB -க்கும் இடையே ஒன்றைத் தேர்ந்தெடுக்க வேண்டும். நாம் எதற்குக் கூடுதல் சாத்தியம் இருக்கின்றதோ அதைத் தேர்ந்தெடுக்க வேண்டும்.

P(VB/To) P(race/VB)

P(NN/TO) P(race/NN)

பின்வருவது சாத்தியங்களாகும் (probabilities)

P(NN/TO) = :021

P(VB/To) = .34

4.4.3 மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள்

மாற்றம் அடிப்படையிலான அடையாளப்படுத்துதல் (Transformation Based Tagging) Bill Tagging என்று அழைக்கப் படுகின்றது. இது பொறி கற்றலுக்கான மாற்றம் அடிப்படையிலான கற்றல் (Transformation based Learning – TBL) அணுகுமுறையாகும். இது விதி அடிப்படையிலிருந்தும் புள்ளியியல் அடிப்படையிலிருந்தும் செய்திகளை எடுக்கின்றது. விதி அடிப்படையிலான அடையாளப்படுத்திகளைப் போல் TBL உம். எந்தச் சொற்களுக்கு எந்த அடையாளத்தைத் தரவேண்டும் என்று குறிப்பிடுகின்றது. ஆனால் அடிப்படையிலான அடையாளப் படுத்திகளைப் போல் TBL ஒரு மொழி கற்றல் நுணுக்கமாகும் (Machine Learning Technique). இதில் தரவிலிருந்து விதிகள் தானாகவே ஊகிக்கப் படுகின்றது. சில HMM- Taggers -ஐ போல் TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் நுணுக்கமாகும் , (Supervised Learning Technique). இது ஏற்கெனவே அடையாளப் படுத்தப்பட்ட பயிற்சி பெற்ற பெருந்தரவை [ore-Tagged Training Corpus] எதிர்ப்பார்கின்றது,ஊகிக்கின்றது. TBL -உக்கு ஒரு குழுமம் அடையாளப்படுத்தும் விதிகள் இருக்கின்றது. முதலில் பெருந்தரவு பெரும்பாலும் பயன்படுத்தப்படும் விரிந்த விதிகளை வைத்து அடையாளப்படுத்தப்படும். பின்னர் சில முன்னதான [original] அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் தேர்ந்தெடுக்கப்படும். பின்னர் குறைவான எண்ணிக்கையிலான அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் பயன்படுத்தப்படும். எவ்வாறு TBL விதிகள்

பயன்படுத்தப்படுகின்றது (How TBL Rules-are applied) என்பது கீழே விளக்கப்பட்டுள்ளது.

Brill's Tagger – ரை கொண்டு பயன்படுத்தப்படும் விதிகள் நடைமுறையைப் பயன்படுத்துவதற்கு முன் அடையாளப்படுத்தி ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும் இந்த கூடுதல் அடையாளங்களை நாம் பெருந்தரவில் பெறலாம்.

Brown Corpus –இல் 'race' என்பதற்குக் கூடுதல் உகந்தது பெயர் என்பதாகும்.

P(NN/race)=.98

P(VB/race)=.02

நாம் முன்னர் பார்த்த race என்பதன் இரு எடுத்துக்காட்டுகளும் NN என்று அடையாளப்படுத்தப்பட்டிருக்கும். முதல் எடுத்துக்காட்டில் NN என்பது தவறான அடையாளம்.

is/VB2 expected/VBN to/To race/NN tomorrow

இரண்டாவது எடுத்துக்காட்டில் race என்பது சரியாக NN என்று அடையாளப்படுத்தப்பட்டிருக்கும்.

The/DT race/NN for/IN outer/JJ space/NN

கூடுதல் உகந்த அடையாளத்தைத் தேர்ந்தெடுத்தப்பின் சேடைடள கூயபபநச அதன் மாற்ற விதிகளை நடைமுறைப் படுத்தும். இது நடக்கையில் race என்பதன் தவறான அடையாளப்படுத்தலில் நடைமுறைப்படுத்தப்படும் விதியைக்

கற்றிருக்கும். முன்வரும் tag to என்றால் NN என்பதை VB என்றும் மாற்றுவதற்து முன் to/To என்று வந்தால் இந்த விதி race/NN என்பதை race/VB என்று மாற்றும்.

Expected/VBN to/To race/NN

Expected/VBN to/To race/VB

எவ்வாறு TBL விதிகள் கற்கப்படுகின்றன என்பது பின்வருமாறு விளக்கப்படும். Brill's TBL வழிமுறை வரைவுக்கு [algorithm] மூன்று முக்கிய நிலைகள் இருக்கின்றன. இது முதலில் ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும். பின்னர் இது ஒவ்வொரு சாத்தியமான மாற்றத்தையும் (Possible Transformation) பரிசோதிக்கும். இறுதியில் இருந்த விதிப்படி தரவை மீள் அடையாளப்படுத்தல் செய்யும். முன்னால் வரும் அடையாளப்படுத்தலை விட குறைந்த மேம்பாடுடைய அடையாளப்படுத்தலான நிறுத்தல் அளவீடு வருவது வரை இந்த மூன்று நிலைகளும் திரும்பித்திரும்பி சொல்லின் சரியான அடையாளம் தெரியும் என்பதை வேண்டுகிறது. அதாவது TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் வழிமுறை வரைவு (Supervised learning algorithm) ஆகும்.

TBL செயல்பாட்டின் வெளிப்யீடு முறைப்படுத்தப்பட்ட மாற்றங்களின் பட்டியலாகும். இவை புதிய பெருந்தரவுக்குப் பயன்படுத்தத்தக்க,நடைமுறை படுத்தத்தக்க அடையாளப்படுத்தல் செயல் முறையை (tagging Procedure) உருவாக்கும். கொள்கை அடிப்படையில் சாத்தியமான மாற்றங்களின் குழுமம் எல்லையற்றதாகும். ஆனால் மாற்றங்கள் ஒவ்வொன்றும் வழிமுறை வரைவைக் கடக்குமாறு மிகப் பொருத்தமான ஒன்றைத் தேர்ந்தெடுக்க TBL ஒவ்வொரு மாற்றத்தையும் எடுத்துக்கொள்ள வேண்டும். இவ்வாறு மாற்றங்களின் குழுமத்தை எல்லைப்படுத்த வழிமுறை வரைவுக்கு ஒரு வழி தேவை. இது சிறிய templates - களின் குழுமத்தை (உருமான மாற்றங்களை) உருவாக்கிச் செய்யப்படுகின்றது.

பின்வருவன Brill's Templetes -களின் குழுமத்தின் பட்டியலாகும். முன்னர் வருகிற (தொடர்நது வருகிற) சொல் Z-ஆக அடையாளப்படுத்தப்படும். இரண்டு சொற்களுக்கு முன்னால் (பின்னர்) வருகிற சொல் Z -ஆக அடையாளப்படுத்தப்படும்.

முன்னர் வருகிற இரு சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற மூன்று சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற சொல் Z -ஆகவும் இரு சொற்களுக்கு முன்னால் (பின்னால்) வரும் சொல் W -ஆகவும் அடையாளப்படுத்தப்படும்.

4.5 சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம்

4.5.1 சொல்வகை அடையாளக் குழுமங்கள் பற்றிய முந்தைய முயற்சிகள்

4.5.1.1 தொடக்ககாலச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம் [1980-க்கு முன்]

1960-களிலும் 1970-களிலும் ஆங்கில மொழிக்கு வேண்டி சொல்வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கும் செயல்பாடு ஐக்கிய அமெரிக்காவில் தொடங்கப்பட்டு விட்டது. ஆரம்பக் காலத்தின் முக்கியமான சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் Klein – Simmons(1963) மற்றும் Gree – Fubin (1971) என்பவர்களால் உருவாக்கப்பட்டது. மற்றொரு முயற்சி பென்சிலியா பல்கலைக்கழகத்தில் மேற்கொள்ளப்பட்டது.

4.5.1.2 ஆங்கில சொல் வகைப்பாட்டுக் குழுமங்கள் (1980-க்குப் பிறகு)

1980-1990-களில் லங்காஸ்டர் பல்கலைக்கழகத்தில் ஆங்கிலத்திற்கு வேண்டி ஒரு தொடர்ச்சியான சொல்வகைப்பாட்டு குழுமங்கள் CLAWS என்ற சொல்வகைப்பாட்டை அடையாளப்படுத்தும் மென்பொருளை பயன்படுத்தி

உருவாக்கப்பட்டது. CLAWS என்ற சொல்வகைப்பாட்டு அடையாளக் குழுமம் LOB தரவுத்தொகுதிக்குப் பயன்படுத்தப்பட்டது. பின்னர் CLAWS -ன் C5 மற்றும் C7 சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் BNC விரிதரவை அடையாளப் படுத்தப் பயன்படுத்தப்பட்டது. இதைத் தொடர்ந்து பல ஆங்கில சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் பயன்படுத்தப்பட்டன. அவையாவன:

1.The TOSCA Scheme
2.The ICE Tagset
3.The PENN Tagset
4.The LUND Tagset
5.The Eng CG Tagset

4.5.2 அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு

ஆங்கிலம் தவிர பிற மொழிகளுக்குச் சொல் வகைப்பாடு அடையாளம் செய்யும் போது சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்தும் தேவை வெளிப்பட்டது. சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் மிக சமீபயத்திய தரமான ஒன்று EAGLES—வழிகாட்டல்களாகும் (The ENGLES-Guidelines) [Leeach and Wilson] இதனுடைய நோக்கம் பலவிதமான திட்டங்களுக்கும் பல மொழிகளுக்கும் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்துவதாகும்.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உள்ளடக்கும் அல்லது உள்ளடக்க வேண்டிய ஒரு குழும பண்புக்கூறுகளைத் தருகிறது. இதனுடன் இந்த எல்லாப் பண்புக்கூறுகளையும் குறியாக்கம் செய்து இடைநிலைப்பட்ட குறியாக்கத்தின் தேவை என்னவென்றால் இரண்டு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கு இடையில் மொழிப்பபெயர்ப்பை

அனுமதிப்பதாகும். இதனால் அவற்றிற்கிடையில் பொருத்தத்தை உறுதி செய்வதாகும். உருப-தொடரியல் அடையாளப்படுத்தலுக்கு EAGLES வழிகாட்டல்களில் மூன்று காட்டல்களில் அமைப்பாக்கம் செய்யப்பட்டுள்ளது.

1. எவை எவற்றைக் கட்டாயமாகக் கருதப்படுகிறது.
2. எவை பரிந்துரைக்கப்படுகின்றது.
3. எவை உருப-தொடரியல் அடையாளப் படுத்தலுக்கு சின்னம் - மதிப்பிணைகளாக வரையறை விளக்கம் செய்யப்படுகிறது. எடுத்துக்காட்டாக 'பால்' என்ற சின்னத்திற்கு EAGLES -இன் பரிந்துரைகளில் ஆண்,பெண் அல்லது பால் மதிப்பிணைகள் படிநிலையாக அமைக்கப்படவோ அமைக்கப்படாமலோ இருக்கலாம்.

EAGLES வழிகாட்டல்களால் எந்தச் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கும் கட்டாயமாக பதிமூன்று பரிந்துரைக்கப்படுகின்றது.

1. Noun (பெயர்)
2. Verb (வினை)
3. Adjective (பெயரடை)
4. Pronoun (பதிலீடு பெயர்)
5. Article (சுட்டிடைச்சொல்)
6. Adposition (உருபு)
7. Adverb (வினையடை)
8. Conjunction (இணைப்பான்கள்)
9. Numeral (எண்:
10. Interjection (வியப்பிடைச்சொல்)
11. Unassigned/unique (இணையற்ற,தனிப்பட்ட)
12. Residual (மீதியான)

13. Punctutation (நிறுத்தற்குறிகள்)

பரிந்துரைக்கப்பட்ட மற்றும் இருப்புச் சின்னங்கள் இந்த முக்கியமான சொல் வகைப்பாடுகளால் ஒழுங்குப்படுத்தப் பட்டுள்ளது. சொல் வகைகளை கடந்து பொருத்தமுறத் தேவையில்லை.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குடும்பத்தைத் திட்டமிடுபவரின் சுதந்திரத்தைக் கட்டுப்படுத்தாமல் அவர் குறிப்பிட்ட எண்ணும் எல்லா அடிப்படையான செய்திகளையும் உள்ளடக்கி நிகழ இயலும் சட்டகத்தைத் தருகின்றது. முன்னர் விளக்கப்பட்ட பல அடையாளக் குழுமங்கள் EAGLES வழிகாட்டல்களில் எளிதில் பொருந்தாத பண்புக் கூறுகளை உள்ளடக்கலாம். இருப்பினும் இவைகள் மற்றொரு அடையாளக் குழுமத்திற்கு மொழிப்பெயர்க்க இயலாத குழுமமும் அவற்றை உள்ளடக்கவில்லை. எனவே அவற்றை EAGLES இடைப்பட்ட அடையாளக் குழுமத்தில் அடக்குவது தேவையில்லாததாகும். இந்தச் சட்டகத்தின் மதிப்பு இது வேறுபட்ட மொழிகளின் மொழியியல் வளங்களின் திண்மையையும் திரும்ப பயன்படுத்தலையும் ஊக்குவிக்கின்றது. மற்றும் சுழலான திரும்பக் கண்டுப்பிடித்தலை ஊக்கமிழக்கச் செய்கிறது.

EAGLES வழிகாட்டலின் முக்கியமான குறை இது. உலக மொழிகளின் மிகக் குறுகிய பகுதியைதான் அடக்குகிறது. ஐரோப்பிய ஒருங்கிணைப்பின் திட்டமாக இது ஆங்கிலம், டச்சு, ஜெர்மனி, டேனிசுஷ், பிரெஞ்சு, ஸ்பேனிசுஷ், போர்த்துகீசு, கிரீக், இட்டாலின் எனற வகைப்பாட்டியல் அடிப்படையில் ஒற்றுமையுள்ள ஒன்பது மொழிகளைத் தான் உள்ளடக்குகிறது. எனவே இவை அல்லாத பிறமொழிகளுக்கு நுஆழுடுநுளு வழிகாட்டல்களின் பயன்பாடு பொருத்தமாகவோ, பொருத்தமற்றதாகவோ இருக்கலாம்.

4.5.3 EAGLES வழிகாட்டல்களின் அடிப்படையில் உருவாக்கப்பட்ட சில சமீபக் காலத்திய சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

EAGLES வழிகாட்டலின் வெளிப்யீட்டிற்குப் பிறகு அவற்றைப் பயன்படுத்தி பலவிதமான மொழிகளுக்குச சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உருவாக்கப்பட்டன. அவை பின் வருமாறு:-

1.The MULTEXT 20 Project

2.The GRACE Project

3.The CRATER Project

EAGLES வழிகாட்டல்களால் அடக்கப்படாத மொழிகளின் சமீப காலத்திய சொல்வகைப்பாட்டு அடையாளக் குழுமங்களில் பின்வருவன அடங்கும்.

1.Tagset Design for Arabic.

2.Tagset Design for Chinese.

3.Tagset Design for Korean.

4.Tagset Design in the Paninijan Tradition.

4.6. தமிழ்த் தரவுத்தொகுதிகளை சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தல்

4.6.1. தமிழ்ச் சொல்வகைப்பாடுகள்

ஒரு மொழியின் சொற்கள் சொல்லியல் அடிப்படையிலும், தொடரியல் அடிப்படையிலும், வகைப்பாடு செய்யப்படலாம். சொற்கள் எந்த இலக்கணக் கூறுக்காகத் திரிபுறும் வாக்கியத்தில் எங்கு, எவ்வாறு இடம் பெறுகின்றன. அவற்றின் செயற்பாடுகள் யாவை என்பதன் அடிப்படையில் அவை ஒரு சொல்

வகுப்பைச் சார்ந்தன என நிர்ணயிக்கலாம். ஆங்கிலத்தில் “Part of speech tag” என்றும்,இதன் படி சொற்களை எட்டு வகுப்புகளாகப் பிரிப்பர்.

1. பெயர் - Noun
2. பெயரடை - Adjective
3. மாற்றுப்பெயர் - Pronoun
4. வினை - Verb
5. வினையடை - Adverb
6. முன்னுருபு - Proposition
7. இணைப்புக்கிளவி - Conjunction
8. வியப்புச்சொல் - Interjection

தற்காலத் தமிழ் இலக்கணத்தை ஆங்கில இலக்கண கண்ணோட்டத்தோடு நோக்கும் இலக்கண நூல்களில் ஆங்கிலச் சொல்வகைப்பாட்டின் தாக்கத்தினைக் காணலாம். பழந்தமிழுக்காக எழுதப்பட்ட மரபிலக்கணங்களில் சொல் வகைகள் தற்காலத் தமிழுக்காக மாற்றியமைக்கப்பட வேண்டும் என மொழியியல் அறிஞர்கள் கருதுகின்றனர். இருப்பினும் தற்காலத் தமிழுக்கு எத்தனை சொல் வகுப்புகள் நிர்ணயிக்கப்பட வேண்டும் என்பதில் அவர்களிடையே ஒருமித்த கருத்து இல்லை.

மரபிலக்கணங்கள் தமிழ்ச் சொற்களை பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், உரிச்சொல் என நான்கு வகுப்புகளாக கொண்டு, பெயரும்,வினையும் முக்கியமானதாகவும் இடையும்,உரியும் துணை வகுப்புகளாகவும் கொள்ளப்படுகின்றது. காலம் காட்டாமல் வேற்றுமை உருபுகளை ஏற்கும் சொற்கள் பெயர்ச் சொற்களாகவும், வேற்றுமை உருபுகளை ஏற்காது காலம் காட்டும் சொற்களை வினைச்சொற்களாகவும் கொண்டுள்ளனர். தனித்து இயங்கும் ஆற்றலின்றிப் பெயர்களையும்,வினைகளையும் இடமாகக்

கொண்டு அவற்றின் முன்னும், பின்னும் வரும் வேற்றுமை உருபுகள், விகுதிகள், சுட்டெழுத்துகள், வினாவெழுத்துக்கள் ஆகியனவும் இடைச்சொல் எனப்பட்டன. பெயர்ச்சொல், வினைச்சொல் போல் தனித்து நின்று பொருளை உணர்த்தாமல் பெயர்ச்சொல்லையோ, வினைச்சொல்லையோ சார்ந்து நின்று அவற்றின் குணத்தை "மிகுதி" என உணர்த்த வல்லன உரிச்சொல்.

மரபிலக்கணத்தார் சொற்பகுப்பை ஆராய்ந்தால் சொற்களைப் பெயர்ச்சொல், வினைச்சொல் என்ற இரு பெரும் பிரிவுகளாகவும், அவற்றைச் சார்ந்து வரும் மூன்றாவது பிரிவுச்சொற்களை இடைச்சொல்,உரிச்சொல் என்று பிரித்தனர். மரபிலக்கணங்கள் உரிச்சொற்களைப் பெயர், வினை இவற்றிற்கு அடையாகக் கொள்கின்றனர். இவ்வாறு மரபிலக்கணங்கள் பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், அடைச்சொல் என நான்கு வகைச் சொல் வகைகளை அடையாளம் காட்டுகின்றன. முற்காலத்தமிழை மரபிலக்கண வழி ஆய்கையில் இந்நான்கு வகை சொல் பகுப்புகள் போதுமானதாக இருந்தது.

தற்காலத் தமிழுக்கான சொல்வகுப்ப்புகள் பற்றி இராசேந்திரன் அவர்கள் தமது கட்டுரையில் விரிவாகப் பேசுகிறார். அவர்கருத்துகள் இங்குத் தொகுக்கப்பட்டுத் தரப்பப்டுகின்றன். தற்காலத் தமிழ் புதிய இலக்கணக் கூறுகளை தன்னகத்தே வளர்த்துக் கொண்டுள்ளமையால் சொற்களை புதிதாக வகைப்படுத்த மொழியியல் அறிஞர்கள் வாதிட்டும், வகைப்படுத்தியும் உள்ளனர். [Asher:1982, லேமன்:1989, Lehman:1989]

லேமன் சொல் வகைகளைப் பற்றி விரிவாக ஆராய்கிறார். [Lehman,1989, 911] தற்காலத் தமிழ் மொழியில் எல்லா அகராதி அல்லது வேர் உருபுகளை நான்கு வகைகளாகப் பிரித்து அவைகளை 1.பெயர் வேர்கள் 2. வினை வேர்கள் 3. வினையடை வேர்கள் 4. பெயரடை வேர்கள் எனக்கொள்கிறார். இவற்றில் பெரும்பாலான வேர்கள் பெயர், வினை வேர்களில் அடங்கும். பெயர்கள் பெயர்

வேர்களையும், வினைகள் வினை வேர்களையும் கொண்டிருக்கும். ஆனால் பின்னுருபுகள் பல வினையடைகள், அளவையடைகள், இணைப்புச் சொற்கள் என்பனவற்றை வடிவ அடிப்படையில் பெயர் அல்லது வினை வேர்களின் திரிபுற்ற அல்லது திரிபுறா வடிவங்களாகக் கொள்ளலாம். பழந்தமிழில் சொல் வகுப்புகள் குறைந்த காரணத்தினால் பல திரிபுற்ற அல்லது திரிபுறாத பெயர், வினை மற்றும் பெயரடை வேர்கள் தற்காலத் தமிழில் பின்னுருபுகள், வினையடைகள், பெயரடைகள், அளவையடைகள் போன்ற பல சொல் வகுப்புகளாக மீள்ளாய்வு செய்யப்பெற்றன. நாம் அறிந்தபடி தற்காலத் தமிழில் இரண்டு சொல் வகுப்புகள் தான் திரிபுகளை ஏற்கவல்லது. அவை பெயரும் வினையும் ஆகும். பெயர்கள் வேற்றுமையுருபு, எண் இவற்றிறகாகத் திரிபுறும். வினைகள் காலம், எண், பால், இடம் மற்றும் சிலவற்றிற்காகவும் திரிபுறும். சொல் வகுப்புகளில் பெயர்கள் சொல்லாக்கத்தை காட்டுகின்றன. பெயர்களை வினைகளிலிருந்து பெறலாம்.

ஆஷர் [1982:101,102] ‘Operational definitions for word classes’ என்ற தலைப்பில் தமிழ்ச் சொற்களை ஆறு வகையாகப் பிரிக்கிறார்: 1. பெயர்ச்சொல் (Noun), 2. மாற்றுப்பெயர் (Pronoun), 3. வினைச்சொல் (Verb), 4. பெயரடை (Adjective), 5. பின்னுருபு (Post position), எண், அளவையடை, 6. இடைச்சொற்கள் (Particles)

இதையே கோதண்டராமன் (1989) தனியே இயங்க வல்ல சொற்களைத தலைமை இலக்கணக் கூறுகளாகக் கொண்டு பத்து வகுப்புகளாகப் பிரிக்கிறார். அவையாவன: 1. பெயர், 2. வினை, 3. பெயரடை, 4. வினையடை, 5. வல்லடை, 6. இணைப்பான், 7. உணர்ச்சி வெடிப்புச்சொல், 8. உணர்வு குறிப்புச்சொல், 9. விளிப்புச்சொல், 10. விளி ஏற்புச்சொல்

மேலும் “விகுதி”, “ஒட்டு” முதலானச் சார்புநிலைக் கிளவிகளைத் துணைமை இலக்கணக் கூறுகளாகக் கொண்டு அவற்றின் “வடிவுக்கும் இயல்புக்கும் ஏற்ப” 1.

விகுதி (Suffix), 2. சொல்லுருபு (postposition), 3. வினைசார் கிளவி (Verb Participle), 4. ஒட்டு (Clitic), 5. சாரியை அல்லது நிரப்பி (Fillers) என ஐந்தாக வகைப் படுத்துகிறார்.

லேமன்(1989:9) உருபனியல் மற்றும் தொடரியல் இலக்கண இயல்புகள் அடிப்படையில் தமிழ்ச்சொற்களை 1. பெயர்கள் (Nouns), 2. வினைகள் (Verbs), 3. பெயரடைகள் (Adjectives), 4. வினையடைகள் (Adverbs), 5. பின்னுருபுகள் (Post position), 6. அளவையடைகள் (qantifiers), 7. வரையறு அடைகள் (Determiners), இணைப்புக் கிளவிகள் (Conjunctions) என எட்டாக சொல் வகுப்புகளைப் பிரிக்கிறார்.

இவ்மூவறிஞர்களும் பெயர், வினை, பெயரடை என்ற சொல் வகுப்புகளைத் தேர்ந்தெடுப்பதில் உடன்படுகிறார்கள். கோதண்டராமனும், லேமனும் வினையடையைச் சொல்வகுப்பாகக் கொண்டு ஆக்ஷரிடமிருந்து வேறுபடுகின்றனர். இணைப்பான் என்று கோதண்டராமனும், இணைப்புக்கிளவி என்று லேமனும் குறிப்பிடும் சொல்வகுப்பை ஆஷர் சொல்வகுப்பாகக் கொள்ளவில்லை. ஆஷரும், லேமனும் பின்னுருபுகளைச் சொல்வகுப்பாகக் கொள்ள கோதண்டராமன் பின்னுருபுகளைச் சொல்லுருபுகள் எனக்குறிப்பிட்டு சார்நிலைக்கிளவியாக வகைப்படுத்துகிறார்.

பெயர்ச்சொல்

ஆஷர் பெயர்ச்சொல் பிரிவை உருபனியல் மற்றும் தொடரியல் அடிப்படையில் சொல் வகுப்பாக நிறுவுகிறார். பெயர்ச்சொற்கள் உருபனியல் அடிப்படையில் வேற்றுமை ஒட்டுகளையும், பன்மை ஒட்டு "கள்" என்பதையும் ஏற்கும். தொடரியல் அடிப்படையில் பெயர்கள் பின்னுருபுத்தொடரின் தலைச் சொல்லாகவும், ஒரு வாக்கியத்தின் எழுவாயாகவோ, செயப்படு பொருளாகவோ

செயல்படும். எழுவாயாக வரும் பெயர்ச்சொல்லின் இடம், எண், பாலுக்கு தகுந்தாற்போல் வினையின் இடம், எண், பால், ஒட்டின் தேர்வும் அமையும். வேற்றுமை உருபுகளையும் ஆக,ஆன கிளவிகளை ஏற்பனவும், எழுவாயாகவும், பயனிலையாகவும் செயல்பட வல்லன பெயர்கள் என்று கோதண்டராமன் குறிப்பிடுகிறார். லேமன் வேற்றுமை பின்னொட்டுகளையும் ஆக,ஆய் பின்னொட்டுகளையும் ஏற்பன பெயர்கள் என்று விளக்குகிறார். மேலும் குறை உருபனில் உள்ள பல பெயர்கள் எல்லா வேற்றுமை பின்னொட்டுக்களையும் ஏற்காது என்பதையும் சுட்டுகிறார். பெயர்ச்சொல் வகுப்பில் மூவருக்கும் அடிப்படை பேதம் இருப்பதாகத் தெரியவில்லை. லேமன் மாற்றுப்பெயர், அளவுப்பெயர், எண்ணுப்பெயர் இவற்றை பெயர்ச்சொல்லின் துணைப்பிரிவாகக் கொள்கிறார்.

மாற்றுப்பெயர்

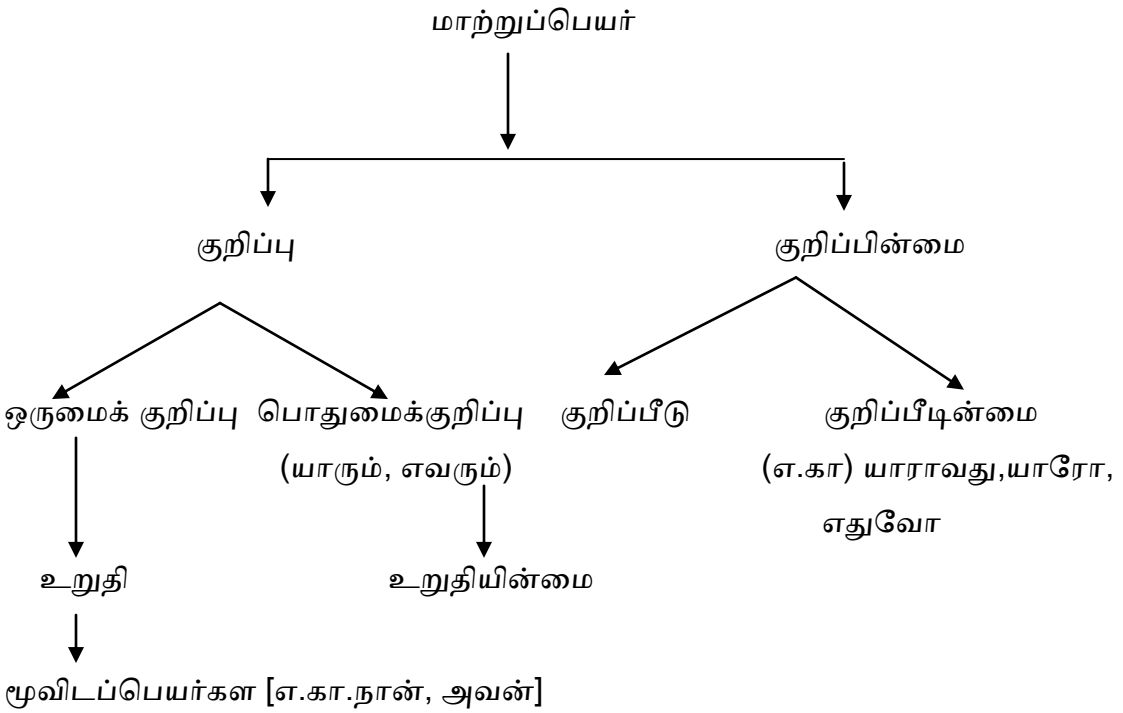
பெயர்ச்சொல்லுக்குப் பதிலாக அல்லது மாறாக வரும் சொல் மாற்றுப்பெயர் எனப்படும். இப்பெயர் முக்கியமாகப் பொருட்குறிப்பு (Signification) அடிப்படையில் பெயரிலிருந்து வேறுப்படும். ஒன்றையோ, பெயர்ச்சொல்லையோ குறிக்கும். ஆனால் மாற்றுப்பெயருக்குத் தக்கவாறு வெவ்வேறு ஆட்களையோ, பொருளையோ குறிக்கும். இதன் காரணமாய் ஒரு மொழியின் பெயர்ச்சொல்லின் எண்ணிக்கை ஏராளமாயிருந்தாலும், மாற்றுப்பெயர்கள் குறைவாகவே உள்ளது.

ஆஷர் மாற்றுப்பெயர்களைத் தனிச்சொல் வகுப்பாகக் கையாள்கிறார். பெயர் சொல்லின் வரையறைகளில் பல இயல்புகளைக் கொண்ட ஒரு நெருங்கிய குழுமத்தைச் சார்ந்திருப்பவை மாற்றுப்பெயர்கள். பெயர்ச்சொற்கள் ஏற்கும் அதே வேற்றுமைப் பின்னொட்டுகளை அவற்றுடன் சேர்க்கலாம். ஆவை பின்னுருபுத் தொடரின் தலைச் சொல்லாக அல்லது எழுவாயகவும், செயப்படுப்பொருளாகவும்

செயற்படலாம். ஒரு வாக்கிய அமைப்பில் வினையின் இடம், எண், பால் ஒட்டைத் தேர்வு செய்யும்.

லேமன் பெயர்த் தொடரில் அடை எடுக்காது தனித்து வரும் பெயர்களை மாற்றுப் பெயர்களாகக் கொள்கிறார். மாற்றுப்பெயர்களைத் தனிநிலை மாற்றுப்பெயர்கள், [எ.கா.நான், அவன்] , ஆக்க மாற்றுப்பெயர்கள் [எ.கா.யாரோ, யாரும்] என்றும் வகைச் செய்கிறார். மூவிடப்பெயர்கள் [Personal Pronoun], சுட்டுப்பெயர்கள் [Demonstrative Pronouns], வினாப்பெயர்கள் [interrogative Pronouns], தன்வயப்பெயர்கள் [Reflexive Pronoun] என பலவகையாகப் பிரிக்கப்படுகிறது. ழுலவகையான மாற்றுப்பெயர்களைக் குறிப்பு (Referential), உறுதி (Definiteness), குறிப்பீடு (Specificity) என்ற பொருண்மைக் கருத்துகள் அடிப்படையில் வனைப்பாடு செய்யலாம்.

லேமன் ,1989:93



மூவிடப்பெயர்கள் காலத்தோடும், இடத்தோடும் தொடர்பு கொண்டிருக்கும். பேசுபவசைத் தன்மை இடம் என்றும், கேட்பவரை முன்னிலை இடம் என்றும், பேசப்படுபவரைப் படர்க்கை இடம் என்றும் செய்யலாம். சுப்ல கூடுதலான பாகுபாடுகளைக் கொண்டு படர்க்கையிடம், தன்மை, முன்னிலை இடங்களிலிருந்து வேறுபடுகின்றது. தன்மை, முன்னிலை, படர்க்கை ஆகிய மூன்றும் அடிப்படையில் ஒருமை, பன்மை என வேறுபடுகின்றது. முன்னிலை இடங்களும், படர்க்கை இடங்களும் தகுதி அடிப்படையில் உயர்வு, உயர்வின்மை என வேறுபடும். படர்க்ளை இடப்பெயர்கள் மட்டுமே பால் அடிப்படையில் ஆண், பெண் என்றும், சுட்டு அடிப்படையில் அண்மை, சேய்மை என வேறுபடும்.

இதற்கான அமைப்பை கீழே அட்டவணை மூலம் காணலாம்

இடம்	ஒருமை	பன்மை
தன்மை	நான்	நாம்,நாங்கள்
முன்னிலை	நீ,நீர்	நீங்கள்
படர்க்கை	அவன்,இவன்,அவள்,இவள், அவர்,இவர்,அது,இது	அவர்கள்,இவர்கள் /அவை,அவைகள் இவை,இவைகள்

தான், நாம், தாங்கள் என்ற தன்வயப்பெயர்கள் நான்காவது இடத்தைச் சாரும். எழுவாயால் வரும் படர்க்கைச் சொல்லுக்கு மாறாய் அல்லது பதிலாய் தன்வயப்பெயர்கள் வரும்.

வினாப்பெயர்களுக்குத் தீர்மானிக்கப்பட்ட குறிப்பு இல்லை. வினாப்பெயர்களைக் குறிப்பீடு பெற்ற (Specified) வினாப்பெயர். குறிப்பீடு பெறாத (non specified) வினாப்பெயர் எனப்பிரிக்கலாம். அவை உயர்திணை, அஃறிணை வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. *யார்*, *என்ன*) குறிப்பீடுப்பெற்ற வினாப்பெயர்கள் படர்க்கையிடம் எண், பால் இவற்றின் வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. எவன், எவள், எவர், எது, எவை) உம் என்ற குறைச்சொல் சேர்ந்த வினாப்பெயர்களுக்கு எடுத்துக்காட்டுகளாகும். குறைச்சொல் ஒ-சேர்ந்த வினாப்பெயர்களான யாரோ, ஏதோ, என்னவோ என்பன குறிப்பீடு பெறாத உறுதியற்ற மாற்றுப்பெயர்களுக்கு எடுத்துக்காட்டுகளாய் அமையும்.

அளவுப்பெயர்

சில, பல, எல்லாம், எல்லோரும் என்ற பெயர்கள் அளவுப்பெயர்கள் பிரிவைச் சாரும். இவை பெயர் அடைகளாகச் செயல்பட வல்லன (எ.கா. சில மனிதர்கள்). பெயர்களுக்குப் பின்னால் வந்து வேற்றுமை ஒட்டுகள் எடுக்க வல்லன (எ.கா. பேனாக்கள் சிலவற்றை வாங்கினேன்). சில, பல என்பவைகளைப் பெயரடைகளாகக் கொண்டு முறையே சிலவை, பலவை என்பனவற்றை மாற்றுப் பெயரொட்டாக்கம் பெற்ற பெயர்களாக ஆயலாம் (எ.கா. சில+அவை=சிலவை)

எண்ணுப்பெயர்

எண்ணுப்பெயர்கள் ஆதார எண் (Cardinal Number), முறைமை எண் (Ordinal number) என இருவனைப்படும். முறைமை எண் ஆதார எண்ணிலிருந்து ஆவது அல்லது ஆம் என்ற குறைச்சொல்லின் (Clitics)சேர்க்கையால் ஆக்கப்படும். ஆதார எண்களிலிருந்து மாற்றுப் பெயரொட்டாக்கம் பெற்ற எண்ணுப்பெயர்களை (Pronominalzed Cardinal Number) பெறலாம். (எ.கா. ஒருவன், ஒருத்தி, ஒருவர்)

வினை

வினையைச் சொல்வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே கருத்து வேறுபாடு இல்லை. வினைகள் காலப்பின்னொட்டுகளையும், இடம் - எண் - பால் பின்னொட்டுகளையும் பிற பின்னொட்டுகளையும் எடுக்க வல்லன் சில வினைகள் குறை உருபனியலை உடையன. அதாவது எல்லா வினைப்பின்னொட்டுகளையும் அவை ஏற்பதில்லை. இணைப்பு வாக்கியங்கள் (Copular Sentences) (எ.கா. அவன் மாணவன்) தவிர்த்த பிற வாக்கியங்களில் கட்டாயமாக வரும் பகுதி வினையாகும். வினைகளை உருபனியல், பொருண்மையியல் இவற்றின் அடிப்படையில் வெவ்வேறு வகையாய் பாகுபாடு செய்யலாம். காலம் மற்றும் பிறவினைப் பின்னொட்டுகளுடன் சேரும்போது ஏற்படும் உருப ஒலியனியல் மாற்றங்கள் அடிப்படையில் வினைகளை மெல்வினை, இடைவினை, வல்வினை என்ற மூன்று வகைகளாகவும் அதற்கு மேற்பட்ட வனைகளாகவும் பிரிக்கலாம். வடிவ மற்றும் செயற்பாடு அடிப்படையில் வினையை முற்றுவினை (finite verb) (எ.கா. வந்தான்) மற்றும் எச்சவினை (non-finite berb) (எ.கா. வந்த, வந்து) என்று பிரிக்கலாம். எச்சவினை பெயரின் அடையாக வருமா அல்லது வினையின் அடையாக வருமா என்பதன் அடிப்படையில் பெயரெச்சம் (Adjectival or relative participle form) (எ.கா. வந்த பையன்) என்றும், வினையெச்சம் (adverbial or verbal participle form) (எ.கா.வந்து போனான்) என்றும் வகைப்படுத்தலாம். தன்வினை (எ.கா.ஆடு) பிறவினை (எ.கா. ஆட்டு) என்ற பாகுபாடு பொருண்மையியல் அடிப்படையில் அமைந்தது. வினையினை தான் செய்யின் தன்வினை என்றும், பிறர் செய்யின் அது பிறவினை என்றும் விளக்கப்பட்ட பொருண்மை வேறுபாடு அவ்வளவு பொருத்தமானதல்ல. (பரமசிவம்,1983:2-3) அடிப்படையிலான வினைகளைத் தனவினைகள் என்றும், (எ.கா. ஓடு, காண்) அவற்றினின்றும் ஒட்டாக்கத்தால் ஆக்கப்பெற்ற வினைகளை (எ.கா. ஓட்டு, காட்டு) பிற வினைகள் என்றும் கொள்ளலாம். வாக்கியத்தில் பயன்படுத்துகையில் செயப்படுபொருளை

எடுக்குமா? என்பதன் அடிப்படையில் வினைகள் செயப்படுபொருள் குன்றியவினை (எ.கா. போ, வா), செயப்படுபொருள் குன்றாவினை (எ.கா. படி, அடி) எனப்பாகுப்படுத்தலாம். பெயருக்கும் வினைக்கும் உள்ள காரக உறவுகளை (Case Relations)கொண்டும், இணைதிறன் (Valency))கொண்டும் வினைகளை வகைப்பாடு செய்யலாம்.

பின்னுருபுகள்

ஆக்ஷர் பின்னுருகைச் சொல்வகுப்புகளில் ஒன்றாகக் கூறுகின்றார். வேற்றுமைப் பின்னொட்டுகளால் திரிபுற்ற பெயர்களுக்குப் பின்வந்து பின்னுருபுத் தொடராக வினையுடன் ஒரு செயற்பாட்டு உறவில் நிற்கும் என்பது சொல் பின்னுருபு என்று தொடரியல் விளக்கம் தரலாம். பின்னுருபுகள் இந்நிலையில் தனித்து நிற்கும் என்பதிலிருந்து கட்டுப்பட்டே நிற்கும் என்பது பலதரப்பட்ட அங்கங்கள் சேர்ந்த சொல்வகுப்பாகும். கோதண்டசாமன்(1989) சாஹ்நிலைக் கிளவி வகை என்ற துணைமைச்சொல் வகையின் கீழ் பின்னுருபைச் சொல்லுருபு எனக் குறிப்பிட்டு வகைபாடு செய்கிறார். பெயர்ச்சொல்லும், வினைச்சொல்லும் காலப்போக்கில் வேற்றுமைப் பின்னொட்டுகள் போல் சொல்லுருபுகளாய் பயன்படுத்தப்படுகின்றன. லேமனும்(1989) பின்னுருபுகளைத் தனிச்சொல் வகுப்பாகக் கொள்கிறார். திரிபுற்ற மற்றும் திரிபுறாத பெயாவடிவுகளும் வினையெச்ச வடிவுகளும் தான் பின்னுருபுகளாகும். எடுத்துக்காட்டாக, தமிழ் பல இடப்பொருள் செயற்பாடுகளைக் குறிப்பிட, இடத்தைக் குறிக்கும் பெயர்களைத் தான் பயன்படுத்துகின்றது. (எ.கா. இல், உள், முன், மேல், கீழ், அடியில்) ஒரு குறிப்பிட்ட சொல்லைப் பெயராகக் கருதுவதா, பின்னுருபாகக் கருதுவதா என்பதில் ஒருமித்த கருத்தில்லை. பின்னுருபுகளாக எடுத்தாளப்படும் பெயர்கள் எல்லாம் உருபனியல் அடிப்படையில் குறை உள்ளவை. அதாவது அவை எல்லா வேற்றுமை உருபுகளையும் எடுப்பதில்லை. பல பின்னுருபுகளாய் வரும் பெயர் வடிவுகள் தொடரியல் அடிப்படையில் குறைப்பாடுள்ளவை. அதாவது அவை

வாக்கியத்தில் பெயர் வரும் எல்லா இடங்களிலும் வருவதில்லை. எல்லாப் பெயர்களைப் போலவும் அவற்றின் இலக்கணச் செயற்பாடுகள் வருவதில்லை. எடுத்துக்காட்டாக, நடு, இடை என்ற பெயர்கள் 'மையம்' என்ற பொருளில் இல் என்ற இடவேற்றுமை பின்னொட்டு எடுத்து நடுவில், இடையில் என பின்னுருபுகளாக மட்டும் வருகின்றன. ஆனால் அவை எழுவாயாகவோ, செயப்படுபொருளாகவோ, பயனிலையாகவோ வருவதில்லை. இப்பெயர் வடிவங்களில் மிகப் பெரும்பான்மையானவையாய் வரும் நெருங்கிய பெயர் மற்றும் வினை வடிவங்களின் குழுமம் பெயர்த் தொடருக்கப் பின்வந்து பின்னுருபுத் தொடராய் அமையும். பின்னுருபாய் வரும் பல பெயர், வினை வடிவங்கள் முறையே தங்கள் பெயர், வினை தொடரியல் இயல்புகளை இழந்து விடுகின்றன. பின்னுருபுகளை அவற்றின் வடிவ மற்றும் அவை எந்தப் பெயர்த் திரிபுக்குப் பிறகு வருகின்றன என்பதில் அடிப்படையில் வனைப்பாடு செய்யலாம். எடுத்துக்காட்டாக, இல் ஒட்டு எடுத்து வரும் நடுவில்,இடையில் என்பன – உக்கு வேற்றுமைக்குப் பின்னர் வரும். (எ.கா. அவர்களுக்கு நடுவில்,இடையில்) வினைவழிப் பின்னுருபுகளை இறந்தகால வினை எச்ச வடிவில் வருவன. (எ.கா. இல்லாமல்,அல்லாமல்) காலங்காட்டா வினையெச்ச வடிவில் வருவன. (எ.கா. தவிர, போல, விட) என வகைப்படுத்தலாம்.

பெயரடைகள்

ஆக்ஷர், லேமன், கோதண்டராமன் ஆகிய மூவருமே பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்கின்றனர். பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே ஒருமித்த கருத்து இல்லை. பெயரடைக்கு ஒரு செயல்முறை வரையறை செய்வது எளிதானதல்ல. பெயரடை தொடரியல் அடிப்படையிலான ஒரு சொல் வகையாகும். பெயரடை, பெயர்த்தொடரில் தலைப்பெயருக்கு முன்னதாக அதன் அடையாக வரும் (எ.கா. பெரிய மரம்), வரையறு அடை (Determiner) பெயரடைக்கு முன் வரலாம். (எ.கா.

இந்தப் பெரிய வீடு) பயனிலை இடத்தில் பெயரடை வருவதானால் அது மாற்றுப்பெயர் ஒட்டாக்கம் பெற்று வர வேண்டும்.(எ.கா. அந்த மாளிகை பெரியது) சொல் வகுப்புகளில் பெயரடை குறைச்சொல் எடுப்பதில்லை. பெயரடைகளைத் தனிநிலைப் பெயரடை (எ.கா.நல்ல, பெரிய, கெட்ட, சிறிய), கூட்டுநிலைப் பெயரடை மற்றும் ஆக்கப்பெயரடை (derived adjective) (எ.கா. அழகான, உயரமான) என வகைப்படுத்தலாம். தனிநிலைப் பெயரடைகள் மரபிலக்கணத்தில் குறிப்புப் பெயரெச்சங்களாகக் கொள்ளப்படுகின்றன. பெயருக்கு பின் வந்து அவற்றைப் பெயரடைகளாகச் செயலாற்றச் செய்யும் ஆன, உள்ள, அற்ற, இல்லாத (எ.கா.அழகான, அழகற்ற, அழகில்லாத) என்ற பின்னுருபுகளைப் பெயரடையாக்கப் பின்னொட்டுகளாகக் கொள்வதா? அல்லது பெயரெச்சங்களாகக் கொள்வதா? என்பது இலக்கணப் பிரச்சனை. பெயரடைகளும் (தெரிநிலை) பெயரெச்சத் தொடர்களும் செயற்பாடு அடிப்படையில் பெயருக்கு அடையாய் வருவன (எ.கா. அந்த நல்ல பையன், அந்த படித்தப் பையன்) ஆனால் (தெரிநிலை) பெயரெச்சங்கள் உடன், பின், பிறகு, போது, முன், மட்டும், வரைக்கும் முதலான உருபுகளுடன் வினையடைத் தொடர்புகளை உருவாக்கும். (எ.கா. வந்தவுடன், வந்தபின், வரும்முன், கூடியமட்டும்) பெயரடைகள் (குறிப்புப் பெயரெச்சங்கள்) இவ்வாறு செயல்படுவதில்லை. (பரமசிவம் (1983:194)) பரமசிவம் மரபிலக்கணங்கள் பெயரெச்சங்கள் எனக் குறிப்பிடும் தெரிநிலைப் பெயரெச்சம் (எ.கா. பாடின பெண்) குறிப்புப் பெயரெச்சம (எ.கா.நல்ல பெண்), எதிர்மறைப் பெயரெச்சம் (எ.கா. படிக்காத பெண்) என்ற மூன்றையும் ஆன ஒட்டாக்காப் பெயரடைகளையும் (எ.கா.அழகான பெண.) பெயரடைகளாகக் கொண்டு அவற்றில் தெரிநிலைப் பெயரெச்சத்தையும், எதிர்மறைப் பெயரெச்சத்தையும் தொடர் என்று குறிப்பிட்டு ஒரு வகையாகவும், குறிப்புப் பெயரெச்சத்தையும் ஆன ஒட்டாக்கப் பெயரடையையும் தனிச்சொல்லாகக் கொண்டு மற்றொரு வகையாகவும் பிரிக்கிறார்.

வினையடைகள்

ஆஷர் (Asher, 1982:101-102) 'Operational Definition for Word Classes' என்ற தலைப்பின் கீழ் சொல் வகுப்புகளில் ஒன்றாக வினையடையைத் தராவிடிலும் சொல்லாக்க உருபனியலைப் பற்றிப் பேசுகையில் அவர் (Asher, 1982:199-205) விடையடைகளின் ஆக்கம் பற்றிக் குறிப்பிடுகிறார். வாக்கியத்தில் அவை இடம் பற்றிக் கூறுகையில் இடவாக்கியங்கள் (Locative Sentences) மற்றும் உள்ளமை வாக்கியங்கள் (existential Sentences) அல்லாத பிற வாக்கியங்களில் வினையடை பொதுவாக எழுவாய் மற்றும் அயல் செயப்படுபொருளைத் தொடர்ந்தோ நேர்ச் செயப்படுபொருளுக்கு முன்னரோ வரும் என்று குறிப்பிடுகிறார். ஒரே வாக்கியத்தில் வேறுபட்ட வினையடைகள் வந்தால் அவை வரும் நிரலைச் சரியாகக் கூறமுடியாது. பொதுவாக, கால வினையடைகள், இட வினையடைகளுக்கு முன்னர் வரும். இடவாக்கியங்கள் இவற்றில் வினையடை எழுவாய் முன்னர் வரும். கோதண்டராமனும், லேமனும் வினையடையைச் சொல் வகுப்பாகக் கொள்கின்றனர். வினையடைகளைத் தனிநிலை வினையடை (Simple adverbs) (எ.கா.நன்கு, உரைக்க) கூட்டுநிலை வினையடை அல்லது ஆக்க வினையடை (எ.கா. அழகாக, அழகாய்) எனப்பகுக்கலாம். திரிபுற்ற அல்லது திரிபுறாத பல பெயர்ச்சொல் மற்றும் வினைச்சொல் வடிவங்களை தொடரியல் அடிப்படையில் நெருங்கிய ஒரு குழு வினையடையாக மீள் ஆய்வு செய்யலாம். பின்னுருபுகளாக வரும் பல சொல் வடிவுகள் வினையடைகளாகவும் வரலாம். (எ.கா. முன்னர், முன்னால், பிறகு) திரிபுறாத ஆனால் குறைச்சொல் ஏ –ஐ விருப்பாக ஏற்றுவரும். பெயர்ச்சொல் வடிவுகள் (எ.கா. அப்பாலே, உள்ளே, அப்புறம்) இடவேற்றுமைக்காகத் திரிபுற்ற பெயர்ச்சொல் வடிவுகள் (எ.கா. பார்த்து, பிந்தி, முந்தி) என்பன வினையடைகளாகவும், பின்னுருபுகளாகவும் செயல்பட வல்லன. பெயரல்லாத வேறுசில சொல்வடிவுகளும் வினையடையாகச் செயல்பட வல்லன. எடுத்துக்காட்டாகச் சுட்டு மற்றும் வினாச் சுட்டுரபுகள் கொண்ட அப்படி, எப்படி, அப்போது, எப்போது, இங்கு, எங்கு, இன்று, என்று

என்பனவற்றையும், பொதுவாக வினையடைகளாகக் கொள்ளலாம். சில இடங்களில் இவை வினையடையல்லாது பெயர்போல் செயற்பாடு செய்யலாம். (எ.கா. இன்று நல்ல நாள்) பல திரிபுற்ற மற்றும் திரிபுறாத பெயர் மற்றும் வினை வடிவங்கள் தொடரியல் அடிப்படையில் வாக்கியநிலை வினையடைகள் அல்லது வினையடை இணைப்பான்கள் என மீள் ஆய்வு செய்யப்படலாம். இவை வாக்கியத்தின் முன்னிடத்தில் வந்து இரண்டு வாக்கியங்களின் பொருண்மை உறவைக் கூறி நிற்கின்றன. எ.கா. அவன் நன்றாய்ப் படித்தான், ஆனால் வெற்றிபெறவில்லை. அப்படியும், ஆனால், இருந்தாலும், எதற்கும் என்பன வாக்கியநிலை வினையடைகளாகும்.

பெயர்ச்சொல்லுடன் ஆகு என்ற வினைச் சொல்லின் காலமிலா வினையெச்ச வடிவான ஆக அல்லது ஆய் சேர்ந்து கிடைக்கும் சொல் வடிவங்களை ஆக்ஷரும், கோதண்டராமனும் வினையடைகளாகக் கொள்கின்றனர். ஆனால் லேமன் ஆக,ஆய் என்பது வினையடையாக்கம் செய்வது மட்டுமன்றி பிற செயற்பாடுகளிலும் வருவதைச் சுட்டிக்காட்டி ஆக,ஆய் என்பதை கட்டுநிலைப் பின்னுருபு அல்லது குறைச்சொல் என்று ஆயலாம் என்கிறார்.

பொருண்மை அடிப்படையில் வினையடைகளைப் பின்வருமாறு பகுக்கலாம்:

1. கால வினையடைகள் (எ.கா.இன்று, நாளை)
2. இட வினையடைகள் (எ.கா.இங்கு, அங்கு)
3. முறைமை வினையடைகள் (எ.கா.மெல்ல, நன்கு)
4. செயல்தொடர் வினையடைகள் (எ.கா.அடிக்கடி, மறுபடியும்)
5. அளவு வினையடைகள்

ஆனால் லேமனும், ஆஷ்ரும் அளவு வினையடைகளை ஒரு தனி வகுப்பாகக் கொள்கின்றனர். பரமசிவம்(1983) வினையெச்சங்களையும் ஒருவகை வனையடையாகக் கொள்கின்றார். வினையெச்சங்கள் காலக்குறிப்பினைக் கொண்டு விளங்குவதால் இவற்றை வினையடைகளாக எல்லோரும் கொள்வதில்லை.

அளவயடைகள்

ஆஷரும், லேமனும் அளவயடைகளைச் சொல் வகுப்பாகக் கொள்கின்றனர். அளவடைகளுக்கு உருபனியல் அடிப்படையில் ஒட்டு மொத்தமான வரையறை தர இயலாது. நெருங்கிய ஒரு குழுமச் சொல்வடிவுகள் (எ.கா. சற்று, முழு, கொஞ்சம், இத்தனை, அத்தனை, இவ்வளவு, அவ்வளவு, எவ்வளவு, நிரம்ப, நிறைய, மிகவும்) பெயர் அடையாக வர வல்லன. ஆனால் இவை வருமிடங்கள் ஒன்றானதல்ல. முழு என்ற அளவடை எப்பொழுதும் தலைப்புப் பெயருக்கு முன் அதனை அடுத்து வரும். (எ.கா.சற்று நேரம்), பெயருடன் ஆக,ஆன, இல்லாத, உடைய சேர்ந்த சொல் வடிவங்களின் முன்னர் மிகவும் வரும். (எ.கா. மிகவும் அழகாக, மிகவும் அழகான, மிகவும் அழகில்லாத) பிற அளவடைகள் பெயரடைரூபெயர் என்ற தொடரின் முன்வரும். (எ.கா.கொஞ்சம்,பெரிய,பாத்திரம்) முழு, இத்தனை, அத்தனை, எத்தனை, மிகவும், தவிர்த்த பிற அளவடைகள் வினையின் அடையாக வினை முன்னிடத்தில் வரும். எ.கா. சற்று மறந்தான், நிறைய சாப்பிட்டான்). மிகவும் என்பன பெயரடையின் அடையாக வந்து பெயரடைத் தொடரில் காணப்படும். (எ.கா கொஞ்சம் சின்னக் கை) கோதண்டராமன் மிகவும் என்பதை பெயர், வினை, பெயரடை, வினையடை ஆகிய நான்கினுக்கு முன் அடையாக வரும். சொற்களை வல்லடை என்ற தனிச்சொல் வகுப்பில் அடக்குகின்றார். (எ.கா. மிகவும் கறுப்பு, மிகவும் நல்ல, மிகவும் வேகமாக, மிகவும் பிடிக்கும்).

வரையறை அடைகள்

பெயருக்குப் பின்னர் வரும் இந்த, அந்த என்ற சுட்டுப்பெயரடைகள் சரையறை அடைகளாகும். ஆவை சுட்டப்படும் பொருள் அல்லது குறிப்புப்பொருள் (referent) பேசுபவர்களிடமிருந்து எவ்வளவு அண்மைப்பட்டு அல்லது சேய்மைபட்டு இருக்கின்றது என்பதைக் குறிக்கவல்லன. (எ.கா இந்தப் பையன், அந்தப் பெண்) கிரியாவின் தற்காலத் தமிழ் அகராதியில் இவை சுட்டுப் பெயரடைகள் என்று குறிப்பிடப்படுள்ளன.

இணைப்பான்கள்

கோதண்டராமனும், லேமனும் இணைப்பான்களைச் சொல்வகுப்பாகக் கொள்கின்றனர். "இரண்டு சொற்றொடர்களை அல்லது சொற்களை இணைப்பதாகவும் ஒரு முழுத்தொடரைத் தன்னோடு இணைப்பதாகவும் அமைந்த சொற்கள் இணைப்பான்கள்" எனக் கோதண்டராமன் விளக்குகின்றார். தமிழில் பொதுவாக குறைச்சொற்கள் (clitics) இருப்பினும் பலவினை வடிவங்களை தொடரியல் அடிப்படையில் இணைப்புச சொற்களாகக் கொள்ளலாம். (எ.கா. ஆனால், அல்லது, இல்லையென்றால்) இணைப்புச் சொற்கள் இருபெயரடைகளையோ, பெயர்களையோ இணைக்கப் பயன்படும். (எ.கா. பழைய ஆனால் நல்ல புத்தகம், ரேக்ஷ்மா இல்லையென்றால் கலா வருவாள். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இணைப்பு இடைச்சொற்கள் என்று குறிப்பிடுகின்றது.

குறைச்சொற்கள்

குறைச்சொற்கள் பற்றி ஆரோக்கியநாதன் (1982) விரிவாக ஆராய்ந்துள்ளார். கோதண்டராமன் குறைச்சொற்களைத துணைமைச்சொல்

வகுப்பாகச் சார்புநிலைக்கிளவி வகையின் கீழ் **'ஒட்டு'** என்று குறிப்பிட்டு வகைப்படுத்துகிறார். அவர் "தொடரில் பல்வேறு இடங்களில் நின்று தொடர்ப் பொருளில் நிலைமாற்றம் செய்யும் மற்றும் பெயர் விகுதியாகவோ, வினைவிகுதியாகவோ" கொள்ள இயலாத தான், உம், ஆ முதலான கிளவிகளை ஒட்டுகள் என்று குறிப்பிடுகின்றார். லேமன் குறைச் சொற்களைப் பற்றிப் பேசும் போது உருபனியல் அடிப்படையில் அல்லாமல் இலக்கணத்தின் ஒலியனியல் விதிகளின்படி சொறகளுள் ஒட்டப்பெறும் கட்டு வடிவங்கள் குறைச்சொற்கள் எனக் குறிப்பிடுகின்றனர். அவை திரிபு அல்லது ஆக்கக் கூறுகளின் பிரதிபலிப்பல்ல. திரிபு அல்லது ஆக்கப் பின்னொட்டுகளைப் போல் அவை ஒரு குறிப்பிட்ட சொல் வகுப்பை மட்டும் சார்ந்து வருவன அல்ல. தமிழில் எல்லாக் குறைச்சொற்களும் பின் குறைச் சொற்களாகும். பெயரடை அல்லது பெயரின் அடையாய் செயல்படும் பெயர்கள் தவிர பிற தொடரியல் வகைப்பாடுகளின் தலைமைச் சொற்களுடன் குறைச் சொற்கள் ஒட்டப்பெறலாம். ஒரு குறிப்பிட்ட ஒலியனியல் வடிவம் கொண்ட குறைச்சொல்லுக்குப் பல பொருண்மைச் செயல்பாடுகள் இருக்கும். உம், ஓ, ஏ, தான், ஆ என்பன குறைச்சொல் வகுப்பின் பாற்படும். ஆக்ஷர் ஏ, தான் ஆகிய தேற்றச் குறிகளையும் (emphatic markers) ஆக என்ற வினாக் குறியையும் (interrogative marker) உம், ஓ ஆகிய இணைப்பான்களையும் பலதரப்பட்ட முக்கியமான வாக்கிய உறுப்புகளுடன் ஒட்டத் தகுந்த மாற்றம் கொள்ளாத கட்டு வடிவங்கள் எனக் குறிப்பிட்டு அவற்றை இடைச்சொற்கள் (Particles) என்ற சொல் வகுப்பில் அடக்குகின்றார். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இடைச்சொற்கள் எனக் குறிப்பிடுகின்றது.

வினைசார் கிளவி

கோதண்டராமன் வந்தபின், வந்தபோது, வந்த உடன் ஆகியவற்றில் காணும் பின், போது, உடன் ஆகிய கிளவிகளை வினைசார் கிளவி என துணைச்சொல் வகையாகச் சார்புநிலைக் கிளவி வகையின் கீழ் தருகின்றார்.

“பெயர்த்தன்மை திரிந்து பெயரெச்சத்தின் பின் விகுதி போல நின்று வினை எச்சங்களை அமைக்கும் கிளவிகளும், வினைசார் கிளவிகள்” என்று குறிப்பிடுகின்றார். செய்யத்தக்க, செய்யக்கூடிய, செய்ய வேண்டிய முதலான கூட்டுநிலைப் பெயரெச்சஆகளில் காணும் தக்க, கூடிய, வேண்டிய என்பனவற்றையும் வினைசார் கிளவிகளாகக் கொள்ளத்தகும் என்கிறார். கிரியாவின் தற்காலத் தமிழகராதி இயற்றையும் இடைச்சொல் என்றே குறிப்பிடுகின்றது. வேற்றுமைப் பின்னொட்டுகளையும் (எ.கா. ஐ, ஆல், கு) அன், ஆன், அள், ஆள் முதலான பால் இட ஒட்டுகளையும் ஆன, ஆக என வரும் பெயரடை, வினையடை விகுதிகளையும் இந்த வகுப்பில் அடக்குகின்றார்.

சாரியை அல்லது நிரப்பி

“தமக்கென்று அகராதிப் பொருளோ, இலக்கணப் பொருளோ இல்லாதனவாய் சொற்கள் ஒன்றோடொன்று சேரும்போது சொற்சேர்க்கைக்கு இசைவாக நின்று சொல் அமைப்புக்கு உதவுவனவாய் அமைந்த கிளவிகள் சாரியை அல்லது நிரப்பி என்று பெயர்பெறும்” என்று கோதண்டராமன் இச்சொல் வகுப்பை வரையறை விளக்கம் செய்கிறார். வீட்டடினை, ஊரினை என்பனவற்றிலுள்ள இனை என்பதையும், வந்தனன், சென்றனன என்பனவற்றிலுள்ள அன் என்பதையும் சாரியை கிளவிக்கு எடுத்துக்காட்டுகளாகக் காட்டுகின்றார்.

மேலே கூறப்பட்ட அனைத்துச் சொல் வகைப்பாடு தமிழுக்குத் தேவையான அடிப்படைச் சொல் வகைப்பாடுகளாகும். இதை தவிர பல நுண்மையான வகைப்பாடுகளையும் உட்படுத்தலாம். எடுத்துக்காட்டாக, வினையை, வினையெச்சம், பெயரெச்சம், வினைமுற்று, துணைவினை என மேலும் வகைப்படுத்தலாம். இச்சொல் வகைப்பாடுகள் சொல்வகை அடையாளக் குழுமத் தேர்வுக்கு அடிப்படையாய் அமையும்.

4.6.2 தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

4.6.2.1 CIIL நிறுவனத்தின் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

இந்திய மொழிகளுக்கு தரவுத்தொகுதிகள் உருவாக்குவதில் முன்னோடியாக விளங்கிய இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கித் தரவுத்தொகுதிகளை அடையாளப் படுத்தியுள்ளது. 2006-இல் இந்தி தரவுத்தொகுதிக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமமாக முடிவு செய்யப்பட்ட சொல் வகைப்பாட்டு குழுமத்தின் பட்டியல் கீழே தரப்பட்டுள்ளது.

CIIL Hindi Corpora Tagset Finalized in August 29.2006

NOUN

1. NN	Commom Noun
2. NNP	Proper Noun (Palce, Person,etc.…)
3. NC	Noun Compound
4. NAB	Noun Abstgract
5. CRD	Cardinal number
6. ORD	Ordinal number

PRONOUN

7. PRP	Personal Pronoun
8. PRI	Indefinite Pronoun
9. PRR	Pronoun Reflexive
10. PRL	Relative Pronoun
11. PDP	Demonstrative Pronoun

VERB

12. VF	Verb Finite Main
13. VNF	Verb non-finite Adjectival/ Verb non-finite Adverbial

14. VAX Verb Auxiliary
15. VNN Gerund/Verb non-finite nominal
16. VINF Verb Infinitive
17. VCC Verb Causative
18. VCD Verb Double Causative

ADJECTIVE

19. ADD Adjective - declinable
20. ADI Adjective - Indeclinable

ADVERB

21. ADV Adverb
22. ADL Adverb of Location

OTHERS

23. QWO Question Words
24. PPO Post position
25. INT Intensifier
26. QTF Quantifier
27. CNJ Conjunction
28. INJ Interjection
27. IND Indiclinable
30. QOT Quotative
31. PRT Particle
32. NEG Negative
33. RDP Reduplication
34. FWD Loan Word [non – nativized words]
35. IDM Idiom
36. PRO Proverb

4.6.2.2 AU-KBC நிறுவனத்தின் தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமம்

AU-KBC நிறுவனம் தமிழுக்கான ஒரு சொல் வகைப்பாட்டுக் குழுமத்தை உருவாக்கியுள்ளது. இது எல்லா இலக்கண மற்றும் சொல்சார் உறுப்புகளையும் உள்ளடக்கிய குழுமம் ஆகும். இது தமிழ்ப்பல்கலைக் கழக மொழியியல் வல்லுநர்கள் உதவியால் AU-KBC -இன் NLP குழுவினரால் 2001-இல் திட்டமிடப்பட்டு உருவாக்கப்பட்டதாகும். பின்வரும் பட்டியல் தமிழின் சொல் வகைப்பாட்டு அடையாளக் குழுமமாகும்.

TAMIL TAGSET

No	Description	TAG	Examples	Suffixes
1	Noun Singular	N.Sg	மரம் (tree)	
2	Noun Plural	N.pl	மரங்கள் (trees) பூக்கள் (flowers)	
3	Pronoun	Pr.Sg	avanY (he), nAnY(I)	
4	Pronoun Oblique from	Pr.Obl.Sg	என்(mine)	
5	Pronoun plural	Pro.Pl	அவர்கள் (they)	
6	Pronoun Plural Oblique	Pro.Obl.Pl	எங்கள் (ours), உங்கள் (yours)	
7	#Euphonic Increment	Eu	பசுவின் (of the cow)	
8	Acusative	Acc	அவனை (him accusative)	ஐ
9	Dative	Dat	அவனுக்கு (to)	உக்கு, கு, க்கு

10	Instumental	Inst	சாவியால் (by the key)	ஆல்
11	Sociative	Soc	அவனோடு (with him), அவனுடன் (with him)	ஓடு, உடன்
12	Locative	Loc	கையில் (in the story)	இல்
13	Ablative	Abl	ஊரிலிருந்து (from city/town	ல்லிருந்து
14	Benefactive	Ben	அவளுக்காக (for her)	உக்காக
15	Genetive	Gen	பசுவின் (of the cow), உடைய (of), அவனது (his)	இன், உடைய, அது
16	Vocative	Voc		அட, இவளே
17	Clitics	Cl		ஆ,ஏ,ஓ,உம்,தான்
18	Verb	V	ஓடு (run)	
19	#Transitive Verb	VT	திருப்பு (turn sth)	
20	#Intransitive Verb	VI	திரும்பு (turn)	
21	Causative	VC	தெரிவi (inform)	
22	Infinitive	Inf	செய்ய (todo)	அ
23	Singular Negative [imperative]	Neg	செய்யாதே (you don't do)	ஆதே
24	Plural Imperative	V.pl	செய்யுங்கள் (you do-honor plu)	உங்கள்

25	Plural Negative [imperative]	Pl..Neg	செய்யாதிருங்கள் (you do-hon or plu)	ஆதிருங்கள்
26	Imperative+Suffixes	Imp	வாட, அம்மா	
27	Present Tense	Pre		கிற், க்கிற், கின்ற், க்கின்ற்
28	Past Tense	Pst		த், ந்த், இன்,த்த், இ, ட், ற்
29	Future Tense	Fut		வ்,ப், ப்ப், உம்
30	First person Singular	1s		என்
31	First person Plural	1pl		ஓம்
32	Second person Singular	2s		ஆய்
33	Second person Plural	2pl		ஈர், ஈர்கள்
34	Third person masculine Singular	3Ms		ஆண்
35	Third person feminine Singular	3 Fs		ஆள்
36	Third person Honorific Singular	3 Hs		ஆர், அர், அனர்
37	Third person Honorific Plural	3 H pl		ஆர்கள்
38	Third person singular Neuter (non-human)	3 N S		awu, wu, um
39	Third person plural	3 Npl		அன, உம்

	Neuter (non-human)			
40	Third person Neuter Future Negative	3NFut Neg	விடாது (it/they), (won't)	ஆ (ஆது?)
41	Future Negative human	Hum Neg	மாட்டேன் (I don't)	மாட்ட்
42	Present/Past Negative (non-future)	Nfut Neg	வரவில்லை (didn't come)	இல்லை
43	#Optative	Opt	வாழ்க (long live)	க, க்க
44	Verbal Participle	Vbp	செய்து (done)	து, ந்து, etc.,
45	Verbal Participle Negative	Vbp Neg	செய்யாமல் (non-done)	ஆமல், ஆது
46	Contitional Positive	Cnd	செய்தால் (if done)	ஆல்
47	Conditional Negative	Cnd Neg	செய்யாவிட்டால் (if not done)	ஆவிட்டால்
48	Adjectval Participle Present/Past	AdjP	செய்த (do pre/pst+sth), செய்கிற (do pre/Pst-th)	த, கிற
49	Adjectival Participle Future	Adjp Fut	ceVyyum (do Future sth)	Um
50	Adjectival Participle Negative	Adjp Neg	செய்யாத (don't do sth)	ஆத
51	Verbal Noun Untensed	Vbn Unt	செய்யல் (doing), செய்தல் (doing), படிக்கல் (reading), படித்தல்l (reading), செய்கை	அல், தல், த்தல், க்கல், கை, க்கை

			(doing), படிக்கை (reading)	
52	Verbal Noun Tensed Positive	Vbn	செய்தது (have done)	ஆது
53	Verbal Noun Tensed Negative	Vbn Neg	செய்யாத்து (have not done)	ஆதது
54	Participal Noun	PpIN	செய்தவன் (one who did)	அன், அள்,து, அர்,அர்கள்.ஐ
55	Adjctival Noun	AdjN	நல்லவன் (good man)	அன், அள்,து, அர், அர்கள்.ஐ
56	Selective (Concessive?)	Sel	அவன், ஆவது(at least him)	ஆவது
57	Ordinal Suffix	Ord	முதலாவது(First)	ஆவது
58	Reportative Suffix	Rep	-	ஆம்
59	Complementizer	Cmp	-	என்று, என்பது, ஆக
60	Auxiliary	Aux	-	-
61	Conjunction	Conj	-	ஆனால், இல்லை என்றால், அல்லது
62	Adjective	Adj	அழகான(beautiful with adj suffix)	ஆன
63	Adverbs	Adv	அழகாக(beautiful with adv Suffix)	ஆக
64	Postposition	PSP	மேல் (above), கீழ் (below)	

65	Quantifiers	Q	ஒன்று (one), சில (few), கொஞ்சம் (some),	
66	Determiners	Det	அந்த (that), இந்த (this)	
67	Wh. Words	INT	என்ன(what), எவை(which), எப்படி(how)	
68	Unknown	UNK	-	-

4.6.2.3 அமிர்தா தொழில்நுட்பக் கல்லூரி பயன்படுத்திய தமிழுக்கானச் சொல்வகை அடையாளக் குழுமம்

அமிர்தா தொழில் நுட்ப கல்வி நிறுவனத்தில் தமிழை மொழிபெயர்ப்பதற்காக நுட்பங்களைப் பயன்படுத்தி அமிர்தா தொழில் நுட்ப கல்லூரி பேராசிரியர்களான தனலெட்சுமி, அனந்த் குமார், சோமன் மற்றும் தமிழ்பல்கலைக்கழக மொழியியல் துறைத்தலைவர் ச. இராஜேந்திரன் போன்றோர்களின் கூட்டு முயற்சியால் உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்திகளின் அட்டவணையைக் கீழேக் காணலாம்.

S.No	POS	DESCRIPTION
1	NN	NOUN
2	NNC	COMPOUND NOUN
3	NNP	PROPER NOUN
4	NNPC	COMPOUND PROPER NOUN

5	ORD	ORDINALS
6	CRD	CARDINALS
7	PRP	PRONOUN
8	PRIN	PRONOUNINTROGATIVE
9	PRID	PRONOUN INDIFINITE
10	ADJ	ADJECTIVES
11	ADV	ADVERB
12	VNAJ	VERB NON-FINITE ADJECTIVES
13	VNAV	VERB NON-FINITE ADVERBS
14	VBG	VERBIAL GERLND
15	VF	VERB FINITE
16	VAX	VERB AUXILARY
17	VNIT	VERB IN FINITE
18	CNJ	CONJUNTION
19	CVB	CONDITIONAL VERB
20	QW	QUESTION WORD
21	COM	COMPLEMENTIZER
22	NNQ	QUANTITY NOUN
23	QTF	QUANTIFIERS
24	PPO	POSTPOSITION
25	DET	DETGERMINERS
26	INT	INTENSIFIER
27	ECH	ECHO WORDS
28	EMP	EMPHASTS

29	COMM	COMMA
30	DOT	DOT
31	QM	QUESTION MARKS
32	RDW	REDUPLICATION

இவ்வட்டவணையைக் கொண்டு உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்தியைக் கொண்டு பல்வேறு அடிப்படையில் உருவாக்கப்பட்ட தொடர்களின் அமைப்பைக் கீழேக் காணலாம்.

பெயர்த்தொடர்

[அந்த <DET> (B-NP) அழகான <ADJ> (I-NP) பறவை <NN> (I-NP)] NP

பெயரடைத்தொடர்

[திரைப்படம் <NN> (B-AJP) சார்ந்த <ADJ> (I-AJP)] AJP

வினையடைத்தொடர்

[அருகே <ADV> (B-AVP)]AVP

இணைப்புச்சொல்

[ஆனால் <CNJ>(B-CJP)] CJP

[என்று<COM> (B-COMP)] COMP

வினைமுற்று தொடர்

[உள்ளது <VF> (B-VFP)] VFP

வினைமுற்றுப்பெறாத வாக்கியம்

[வெளி வந்த (VNAJ) (B-VNP)] VNP செய்திக் <NNC> < B-NP> குறிப்பு <I – NP> <NNC>

[விரைந் து <VNAV>(B-VNP)] VNP செய்தான் <VF>

gerundial chunk

தொழிற்சாலை <NN> [அமைப்பதில் <VBG>(B-VGP)] VGP தாமதம் <NN>

ஒரு விரிதரவில் எவ்வாறு சொற்களை சொல்வகை அடையாளப்படுத்தலாம் என்பதற்கான ஒரு எடுத்துக்காட்டைப் பார்ப்போம்.

வளாகத் <NNC> <B-NP>
தேர்வில் <NNC> <I-NP>
வேலைவாய்ப்பு <NN> <B-NP>
பெற்ற <VNAJ> <B-VNP>
மாணவர்களின் <NN> <B-NP>
பட்டியல் <NN> <I-NP>
வெளியிடும் <VNAJ> <B-VNP>
விழா <NN> <B-NP>
திங்கள்கிழமை <NNP> <B-NP>
நடைப்பெற்றது<VF> <B-VFP>
. <DOT> <O>

4.4.2.4 IIIT ஹைதராபாத்தில் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக் குழுமம்

இந்திய மொழிகளுக்கிடையிலான மொழிபெயர்ப்பிலும், ஆங்கில மொழியிலிருந்து இந்திய இயந்திர மொழிப்பெயர்ப்பிலும் சூடுப்பட்டுள்ள ஷைதராபாத்தில் உள்ள IIIT நிறுவனம் இந்தியாவில் இயற்கை மொழி ஆய்வுக்கு முன்னோடியாக உள்ளது. இந்நிறுவனம் உருவாக்கிய இந்திய மொழிகளுக்கான (குறிப்பாக இந்திக்கான) சொல்வகைப்பாட்டு அடையாளக் குழுமம் பரவலாகப் பயன்படுத்தப்படுகின்றது. இதன் பட்டடியல் கீழே தரப்பட்டுள்ளது.

NN Common Noun

NNP	Proper Noun
*C	for all Compounds
QC	Cardinal No
QO	Ordinal No
PRP	Pronoun
VF	Verb finite main
VNF	Verb Non-finite adverb
VAUX	Verb Auxiliary
VNN	Gerund/Verb non-finite nominal
VINF	Verb Infinitive
JJ	Adjectives
RB	Adverbs
PSP	Post position
QF	Quantifiers
WQ	Question words
CC	Conjunction
NEG	Negative
INT	Interjection
CL	Classifier
SVM	Special

4.6.2.5 AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு

AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீட்டுப் பட்டியல்கள் கீழே தரப்பட்டுள்ளது. இப்பட்டியல் நிறுவன அறிவியலாளர் எல். சோபா என்பவரால் உருவாக்கப்பட்டது.

NOUNS

AU-KBC	CIIL	IIT-H
	NN - Common Noun	NN - Common Noun
	NNP - Proper Noun	NNP - Proper Noun
	NC - Noun Compound	*C - for all Compound
	CRD - Cardinal No	QC - Cardinal No
Ord - Ordinal No	ORD - Ordinal No	QO - Ordinal No
N.Sg - Noun Sg		
N.Pl - Noun Pl		
	PRP – Personal pronoun	PRP –Pronoun
	PRI – Idenfinite Pronoun	
	PRR – Reflexive Pronoun	
Pr.sg – Pronoun Sg		
Pr.Pl – Pronoun Pl		
Pr.Obl.Sg – Pronoun Oblique Sg		
V – Verb	VF – Verb Finite Main	VF - Verb Finite Main
	VNF – Ver non finite Adverbial adjectival	VNF – Ver non finite Adverbial adjectival
Vbn – verbal noun untensed, Vbn – Vbn Tensed Positive, Vbn Neg – Vbn Tensed Negative	VNN – Gerund/Verb non-finite nominal	VNN – Gerund/Verb non-finite nominal
Inf – Infinitive	VINF – Verb Infinitive	VINF – Verb Infinitive

VC – Verb causative	VCC – Verb Causative	
	VCD – Verb Double Cousative	
VT – Verb Transitive		
VI – Verb Intransitive		

ADJECTIVES

AU-KBC	CIIL	IIIT-H
Adj – Adjective		JJ – Adjective
	ADD – Adjective Declinable ADI – Adjective Indeclinable	

ADVERBS

AU-KBC	CIIL	IIIT-H
Adv – Adverb	ADV – Adverb ADL – Adverb of Location	RB – Adverb

OTHERS

AU-KBC	CIIL	IIIT-H
PSP – Postposition	PPO –Post position	PSP – Post position
Q – Quantifiers	QTF – Quantifiers	QF – Quantifiers
CNJ – Conjunction	CNJ - Conjunction	CC – Conjunction
	INT – Intensifer	INT – Intensifer

	NEG – Negative	NEG – Negative
	INJ - Interjunction	INJ – Interjunction
	PRT – Particle	RP – Particle

OTHERS

CIIL
IND – Indeclinable
QOT – Quotative
RDP – Reduplication
LWD – Loan Word
IDM – Idion
PRO – Proverb

இங்கு சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் பற்றிய முந்தய முயற்சிகள், சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பாக விளங்குகிறது. அவை CIIL சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் AUKBC நிறுவனத்தின் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக்குழுமம் AUKBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு குறிக்கப்பட்டுள்ளது. இவைகளை தேவை அடிப்படையில் பயன்படுத்த தேர்ந்தெடுத்துக் கொள்ளலாம்.

4.6.2.6 பிற சொல்வகை அடையாளப்படுத்திகள்

தமிழுக்கு வேண்டி பல சொல்வகை அடையாளப்படுத்திகள் உருவாக்கப்பட்டுள்ளன. அவற்றில் சில முக்கியமானவை கீழே தரப்பட்டுள்ளன.

4.6.2.6.1 வாசு அரங்கநாதனின் டேக் தமிழ்

வாசு அரங்கநாதனால் உருவாக்கப்பட்ட டேக் தமிழ்ச் (Vasu Ranganathan's Tag Tamil) சொல்-ஒலியனியல் (Lexical-phonology) அணுகுமுறை அடிப்படையில் உருவாக்கப்பட்டது. டேக் தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வை உருபன்களின் வருகை முறையை வரிசை அட்டவணையைப் பயன்படுத்தி செய்கினறது. டேக் தமிழ் சொல் வகை அடையாளப்படுத்தலையும் உருவாக்கலையும் செய்கிறது.

4.6.2.6.2 கணேசனின் சொல்வகை அடையாளப்படுத்தி

கணேசன் தமிழுக்கு வேண்டி ஒரு சொல் வகை அடையாளப்படுத்தியை உருவாக்கியுள்ளார். அவருடைய அடையாளப்படுத்தி இந்திய மொழிகளின் மைய நிறுவன தமிழ் விரிதரவில் நன்கு வேலை செய்கின்றது. பிற தரவுத்தொகுதிகளின் இதன் திறன் மதிப்பீடு செய்யப்பட வேண்டும். இவர் தமிழுக்கு வேண்டி விரிவான சொல்வகை அடையாளப்படுத்திகளின் குழுமத்தை வைத்திருக்கின்றார். இவர் இந்திய மொழிகளின் மைய நிறுவன தரவுத்தொகுதியின் ஒரு பகுதியை அகராதியாலும் உருபனியல் பகுப்பாய்வியாலும் சொல் வகைக்காக அடையாளப்படுத்தி உள்ளார். பின்னர் இவை அடையாளப்படுத்தப்பட்ட விரிதரவை மனிதசக்தியால் திருத்தி மீதியுள்ள பெருந்தரவுக்கு பயிற்சி தந்துள்ளார். இவருடைய சொல்வகை அடையாளப்படுத்தியில் சொல்வகை அடையாளப்படுத்தல் ஒவ்வொரு உருபன்களுக்கும் தரப்படுகின்றது.

4.6.2.6.3 RCILTS-இன் தமிழ் கதம்பம்

கதம்பம் ஆவணத்தில் தரப்பட்டுள்ள சொல்லுக்குச் சொல்வகை அடையாளத்தைத் தருகிறது. இது தமிழ்மொழியியல் அடிப்படையிலான ஊக விதிகளைப் பயன்படுத்துகின்றது. அகராதியையோ உருபனியல் பகுப்பாய்வியையோ பயன்படுத்துவதில்லை. இது பெரிய ஆவணங்களில் 80% திறனைத் தருகிறது. இது பன்னிரண்டு ஊக விதிகளைப் பயன்படுத்துகிறது. இது இட-எண்-பால், காலம் மற்றும் வேற்றுமை குறியீடுகளின் அடிப்படையில் சொல்வகை அடையாளங்களை அடையாளம் காண்கிறது. துனியாக நிற்கும் சொற்கள் சொல்வகை அடையாளப்படுத்தியில் சேகரிக்கப்பட்ட பட்டியலில் இருந்து பரிசோதிக்கப்படுகிறது. இது அறியப்படாத சொற்களுக்கு (unknown words) நிரப்பு விதியைப் பயன்படுத்துகிறது. இது bபைசயஅ மாதிரிகளைப் பயன்படுத்தி முன்வரும் சொல்லின் சொல்வகைப்பாட்டைக் கொண்டு அறியப்படாத சொல்லை அடையாளம் காண்கிறது.

4.7. இந்தி மொழியிலிருந்து இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையில் சொல்வகைப்பாடு அடையாளப்படுத்தல்

தமிழில் முக்கியமான சொல்வகைப்பாடுகள் பெயர், வினை, பெயர்த்தொடர், வினைத்தொடர் என்பவைகளாகும். இந்தியமொழியிலிருந்து இந்தியமொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையில் (இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறை) 21 சொல்வகைப்பாடுக் குழுமங்கள் பயன்படுத்தப்படுகின்றன. சொல்வகைப்பாட்டு அடையாளப்படுத்தி சொல்வகைப்பாட்டு அடையாளப்படுத்தும் இயந்திரத்தால் ஆதரவளிக்கப்படுகின்றது.

4.7.1. இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறையின் சொல்வைப்பாட்டுக் குழுமம்

இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறையில் உட்படுத்தப்பட்டுள்ள எல்லா மொழிகளுக்கும் ஒரு பொதுவான சொல்வகைப்பாட்டு அடையாளக் குழுமம் பயன்படுத்தப்படுகின்றது. இச்சொல்வகைப்பாடுக் குழுமம் தான் தமிழுக்கும் பயன்படுகின்றது. சொல்வகைப்பட்டு அடையாளக் குழுமத்தில் பயன்படுத்தப்பட்டுள்ள அடையாளங்கள் கீளே விளக்கப்பட்டுள்ளன.

1. NN (Noun, பெயர்)

NN என்ற அடையாளம் இலக்கணத் தகவல்கள் அடிப்படையில் வேறுபடுத்தப்படாமல் பொதுவான பெயர்களுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா. நல்ல <JJ> கோவில் <NN>

2. NST (Noun denoting spial and temporal expressions, இட மற்றும் காலத்தைச் சார்ந்த பெயர்கள்)

முன்னால், *பின்னால்*, *அப்புறம்* போன்ற இட மற்றும் காலம் சார்ந்த சொற்கள் NST எனஅடையாளப் பயன்படுத்தப்படும்.

கோவிலின் <NN>முன்னால் <NST> நிற்கிற ஆள்

3. NNP (Pronoun, இயற்பெயர்)

எல்லா இயற்பெயர்களும் NNP என அடையாளப்படுத்தப்படும்.

4. PRP (Pronoun, மாற்றுப்பெயர்)

எல்லா மாற்றுப்பெயர்களும் PRP என அடையாளப்படுத்தப்படும்.

சென்னையிலிருந்து <PRP> மகாபலிபுரத்திற்கு <PRP> இருபது <QC> கிலோமீட்டர்கள் <NN>.

5. DEM (Demenostrative, சுட்டு)

எல்லாச் சுட்டுக்களும் DEM என அடையாளப்படுத்தப்படும்.

அந்தக் <DEM> கோவில் <NN>

6. VM (Verb main, முதன்மை வினை)

எல்லா முதன்மை வினையளும் VM என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சுற்றுலா <NN> சென்றனர்< VM. <SYM>

7. VAUX (Auxiliary Verb, துணைவினை)

எல்லாத் துணைவினைகளும் VAUX என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சென்னையிலிருந்து <NNP> மகாபலிபுரம் <NNP> செல்ல <VM> புறப்பட்டுக் <VM> கொண்டிருந்தனர் <VAUX>. <SYM>

8. JJ (Adjecive, பெயரடை)

பெரிய <JJ> கோவில் <NN>

9. RB (Adverb, வினையடை)

எல்லா முறைமை வினையடைகளும் RB என அடையாளப்படுத்தப்படும்.

பஸ் <NN> வேகமாகச் <RB> சென்று <MV>கொண்டிருந்தது <VAUX>. <SYM>

10. PSP (Postposition, பின்னுருபு)

எல்லா பின்னுருபுகளும் PSP என அடையாளப்படுத்தப்படும்.

அவர்களுடன் <PRP> கூடவே <PSP> வழிகாட்டியும் <NN> சென்றார் <VM>. <SYM>

11. CC (Conjucts, both co-ordinative and subordinative, இணைப்புகள், சமநிலை இணைப்பும் துணைநிலை இணப்பும்)

சமநிலை இணைப்புகளும் துணைநிலை இணைப்புகளும் CC என அடையாளப்படுத்தப்படும்.

கோயில்கள் <NN> மற்றும் <NN> சர்ச்சுகள் <NN> வழியில் <NN> இருந்தன <VM>.

12. WQ (Question Words, கேள்விச் சொற்கள்)

என்ன, எங்கே, எப்பொழுது போன்ற கேள்விச் சொற்கள் WQ என அடையாளப்படுத்தப்படும்.

13. QF (Quatifiers)

எல்லா அளவை அடைகளும் QF என அடையாளப்படுத்தப்படும்.

பல <QF> கோவில்களைக் <NN> கண்டு <VM>மகிழ்ந்தனர் <VM>.

14. QC (Cardinal, அடிப்படை எண்)

எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

ஒன்று, இரண்டு, பத்து, நூறு போன்ற எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

இரண்டு <QC> கிலோமீட்டர் <NN> தூரத்தில் <NN>

15. QO (Codinal, முறைமை எண்)

எல்லா முறைமை எண்களும் QO என அடையாளப்படுத்தப்படும்.

சென்னையில் <NNP> காண <VM> வேண்டிய <VAUX> இரண்டாவது <QO>

இடம் <NN> மெரீனா <NNP> கடற்கரை <NN>ஆகும் <VAUX>. <SYM>

16. INFT (Inensifier, மிகையடை)

மிகையடை மொழியின் பெயரடைகளையுடனும் வினையடைகளுடனும் பயன்படுத்தப்படுகின்றது. எல்லா மிகையடைகளும் INFT என அடையாளப்படுத்தப்படும்.

மிகப் <INTF> பெரிய <JJ> கோபுரம் <NN>

பஸ் <NN> மிக <JJ> வேகமாகச் <RB> சென்றது <VM>. <SYM>

17. NEG (Negation, எதிர்மறை)

இல்லை போன்ற எதிர்மறைச் சொற்கள் NEG என அடையாளப்படுத்தப்படும்.

அங்கு <NST> குளிப்பதற்கு <VM> இடம்<NN> இல்லை <NEG>. <SYM>

18. SYM (Speical Symobl, சிறப்புக் குறியீடு)

பிற அடையாயாளங்களால் பாகுபடுத்தப்பட இயலாத சொற்கள் SYM என் அடையாளப்படுத்தப்படும். $ போன்ற சிறப்புக் குறியீடுகள் SYM என அடையாளப்படுத்தப்படும்.

19. RDP (Reduplication, இரட்டிப்பு)

இந்திய மொழிகளில் சில சொற்கள் இரண்டு தடவை எழுதப்படுகின்றன. அம்மாதிரியான சொற்கள் RDP என அடையாளப்படுத்தப்படும்.

யாத்திரிகள் <NN> மலையில் <NN> மெல்ல மெல்ல <RDP> அடிவைத்து <MV> ஏறினார்கள் <VM>. <SYM>

20. ECH (Echo Words, எதிரொலிச் சொல்)

எதிரொலிச் சொற்கள் இந்திய மொழிகளில் பொதுவாகக் காணப்படும்.

அவர்கள் <PRN> தடதட <ECH> என்று படிகளில் <NN> ஏறினார்கள் <VM>.

21. UNK (Unkown Words, தெரியாத சொற்கள்)

தெரியாத சொற்கள் UNK என அடையாளப்படுத்தப்படும்.

4.7.2 சொல்வகைப்பாடு அடையாளப்படுத்தும் பொறியின் செயல்பாட்டுத் தனிக் குறிப்பீடுகள் (Functional Specifications of POS tagging Engine)

வாக்கியங்களில் வரும் ஒவ்வொரு சொல்லுக்கும் சொல்வகை அடையாளப்படுத்தும் செயல்பாடு சொல்வகை அடையாளப்படுத்தல் (Parts of Speech (POS) tagging) எனப்படும். வாக்கியங்களின் வரும் உறுப்புகளின் பங்களிப்பை ஆய்வதற்குப் பெயர், வினை, பெயரடை, வினையடை போன்ற சொல்வகைப்பாடுகளைக் கண்டுபிடிப்பது உதவும். இதற்கு விதி அடிப்படையிலானது (rule-based), புள்ளியியல் அடிப்படையிலானது (statistics based), மாற்றம் அடிப்படையிலானது (transformation based) எனப் பல அணுகு முறைகள் இருக்கின்றன. இ-இ இயந்திர மொழிபெயர்ப்பில் புள்ளியியல் அணுகு முறை பயன்படுத்தப்படுகின்றது.

4.7.2.1 இடுவரல் விடுவரல் தனிக்குறிப்பீடுகள்

Input TKN_

Output CAT_

CAT_ என்ற தனிதன்மையின் எல்லாச் சாத்தியங்களான மதிப்புகளும் இங்கு பட்டியலிடப்படவேண்டும். NN, PRP, VM மற்றும் பிற.

இடுவரல் தனிக்குறிப்பீடுகள் TKN_ என்ற தனிதன்மை இடுவரல் சொல்வகைப்பாடு அடையாளப்படுத்திக்குத் தரப்படும் SSF-இல் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும்.

விடுவரல் தனிக்குறிப்பீடுகள் CAT_ என்ற தனித்தன்மை சொல்வகைப்பாடு அடையாளப்படுத்தியால் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும். எனவே விடுவரல் SSF சில செல்லுபடியாகும் மதிப்புகளுக்கு CAT தனிப்பண்பைக் கொண்டிருக்கவேண்டும்.

CAT_ என்பது PRP, NN போன்ற மதிப்புகளைக் கொண்டிருக்கும்.

எல்லா CAT_ மதிப்புகளையும் இங்கு பட்டியலிடவும்; பின்னர் ஒவ்வொன்றிற்கும் ஒரு சுருக்கமான வருணனையைத் தரவும்.

PRP - Pronoun

NN – Noun

எடுத்துக்காட்டு:

Input

ADDR_	TKN_	OTHR
1	Afwa	<fs af =' afwa pronn…'>
2	pEyanE	<fs af =' pEyanE, noun…'>
3	Pola	<fs af =' pola, psp…'
4	irukkirYAn	<fs af =' irukkiRaan, verb…'

Output

ADDR_	TKN_	CAT_	OTHR
1	afwa	PRP	<fs af =' afwa, pron…'
2	pEyanE	NN	<fs af =' pEyanE, noun…'
3	pola	PSP	<fs af =' pola, psp…'>
4	irukkirYAn	Verb	<fs af =' irukkirYAn, verb…'

(குறிப்பு: நீங்கள் ஒரு பண்புக்கூறை (feature) பயன்படுத்துவதானால் இடுவரல் மாதிரியிலும் விடுவரல் மாதிரியிலும் காட்டவும்.)

4.7.2.2 சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம்

சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம் ஒரு தொகுதியிலிருந்து மற்றொரு தொகுதிக்கு வழியமைப்புக் கட்டுப்பாட்டு ஒழுக்கின் (program control flow) முழுக் காட்சியையும் காட்டும். இது வழியமைப்புக்குள் பல்வேறுபட்ட தீர்மான இடங்களைக் காட்டும்.

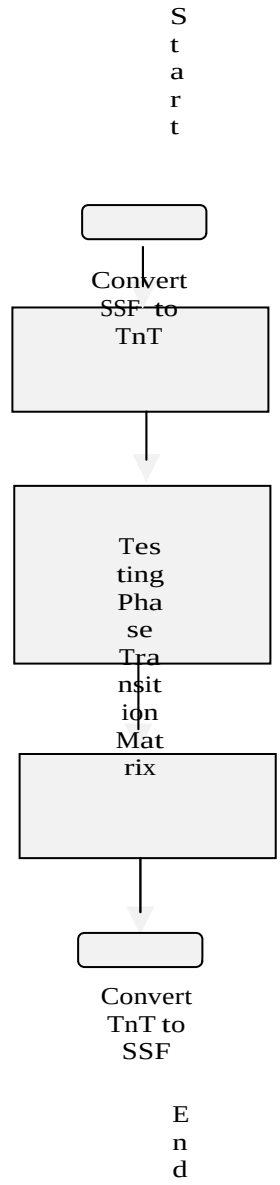


Fig 28 — Flow chart for POS Tagger

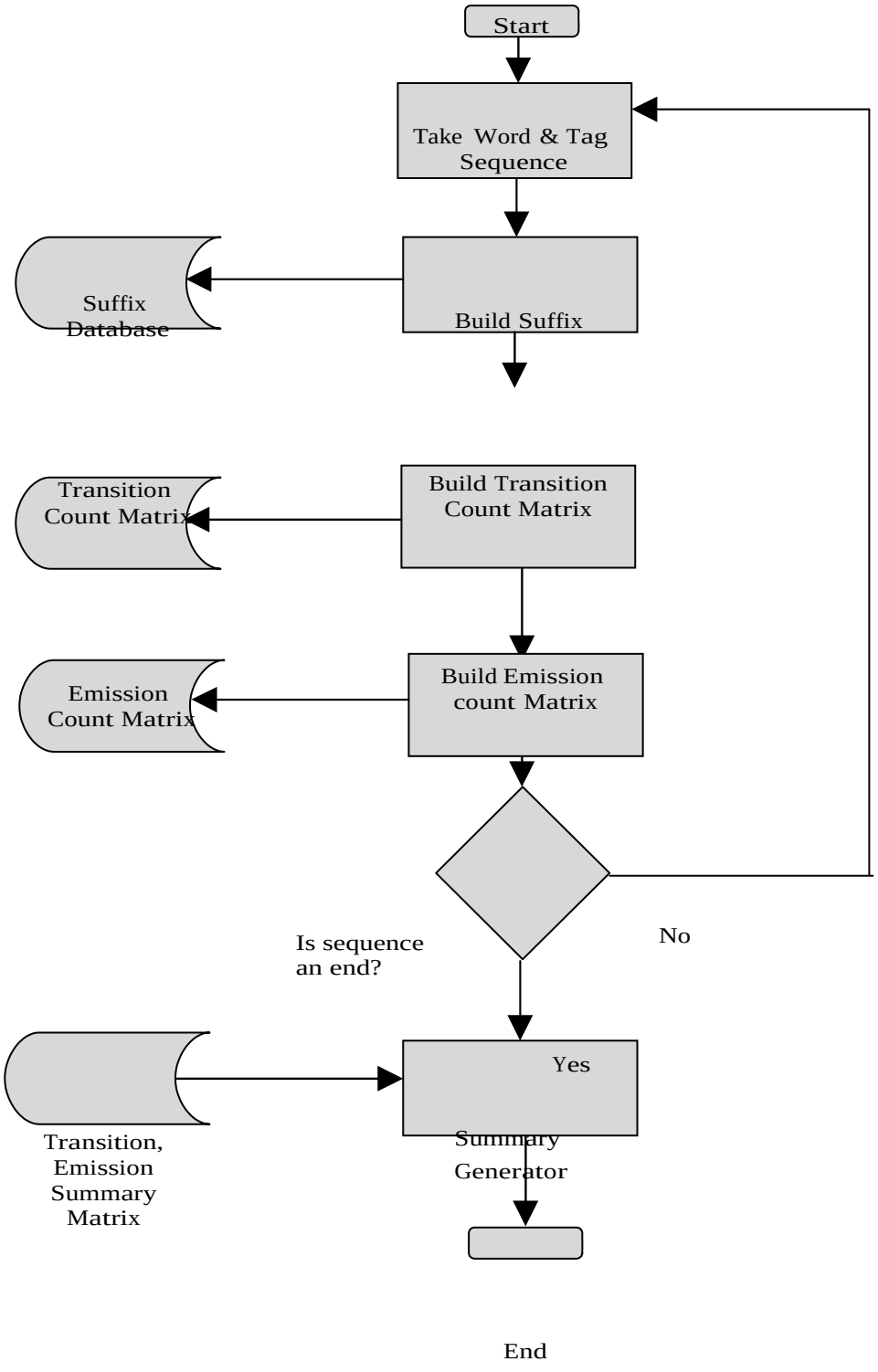


Fig.29 — Flow Chart of Training Phase

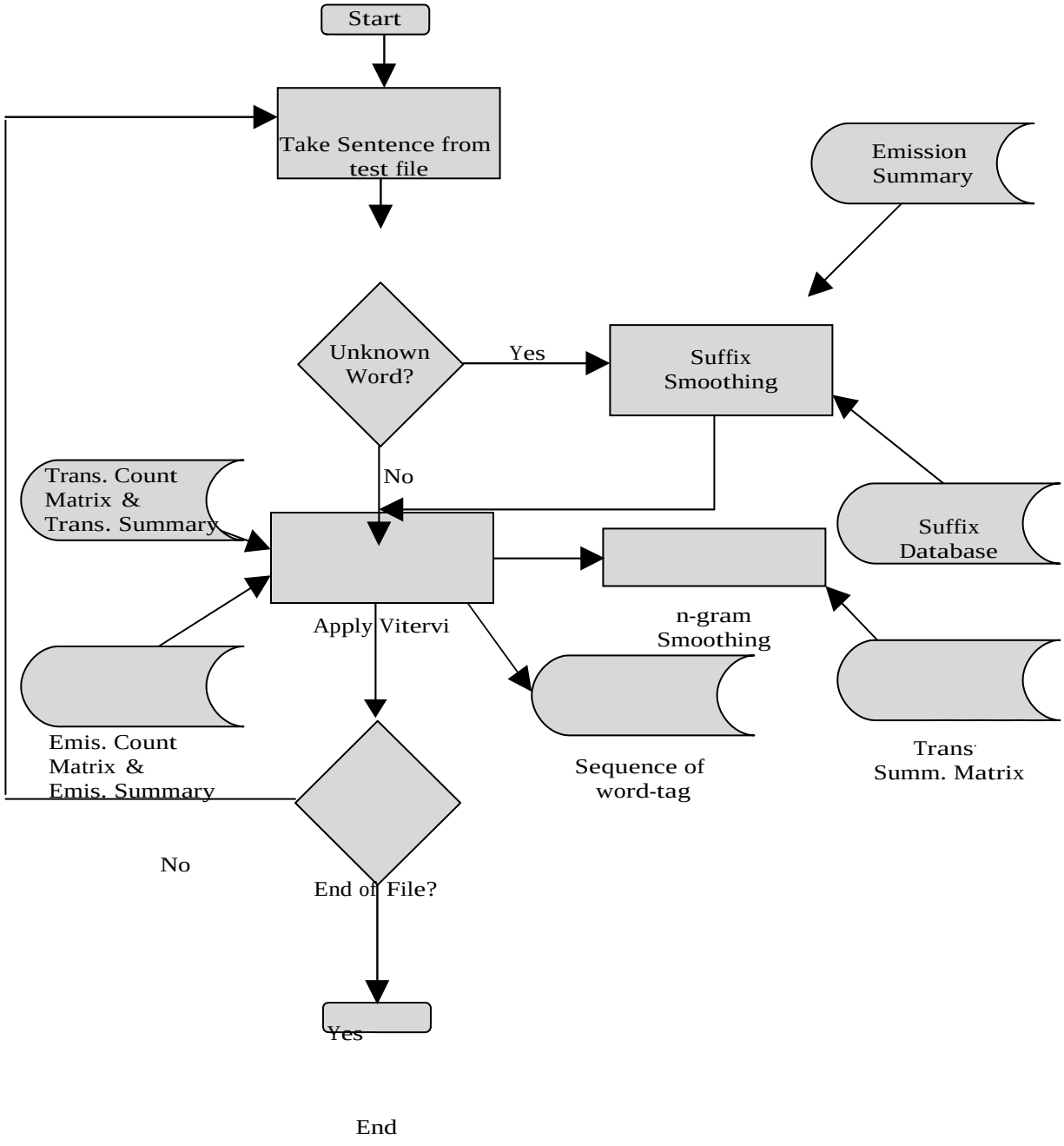


Fig.30 — Flow Chart of Testing Phase

4.7.2.3 செயல்பாட்டு வருணனைகள்

1) SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று (Covert SSF to TnT Format)

   பயிற்சிக் கோப்பை SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று. மேலும் சோதனை கோப்பையும் வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று.

2) TnT வடிவமைப்பிலிருந்து SSF வடிவமைப்புக்கு மாற்று (Convert TnT Format to SSF)

   TnT வடிவமைப்பில் உள்ள சொல்வகைப்பாடு அடையாளப்படுத்தியால் உருவாக்கப்பட்ட இடுவரலை SSF வடிவமைப்புக்கு மாற்று.

3) எழுத்துக் கிளையை உருவாக்கு (Build Letter Tree)

   முன்னொட்டு கிளைத் தரவு அமைப்பை உருவாக்கப் பயிற்சிக் கோப்பிலிருந்து சொற்களையும் அடையாளங்களையும் எடுக்கவும். கிளையமைப்பில் சொல்லையும் அடையாள நிகழ்வெண்ணையும் சேகரிக்கவும். சொல்லையும் அதன் நிகழ்வெண்ணையும் பங்கெடுப்பாளர்களாகக் எடுத்துகொண்டு எழுத்துக் கிளையை உருவாக்கவும்.

4) புடைபெயர்வு எண்ணிக்கை சட்டத்தைக் உருவாக்கவும் மற்றும் வெளிவரு எண்ணிக்கை சட்டத்தையும் உருவாக்கவும் (Build Transtition Count Matrix and Build Emission count matrix)

   அடையாள வரிசை மற்றும் அதன் நிகழ்வெண்ணின் கலவையை உருவாக்கவும்.

5) மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடித்தல் (Finding Best Tag Sequence)

   ஒரு வாக்கியத்தை எடுக்கவும் மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடிக்க Veterbi algorithm-ஐப் பயன்படுத்தவும்.

6) அடையாள வரிசையைக்கு மெருகிடுதல் (Smoothing for Tag sequence)
   ஒரு அடையாள வரிசை இல்லாவிட்டால் சொல்வகைப்பாட்டின் பங்கெடுப்பாளர் அடிப்படையில் மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்.
7) சொற்களுக்கு மெருகிடுதல் (Soothing for Words)
   சட்டகங்களில் சொல் கண்டுபிடிக்கப்படாவிட்டால் பின்னொட்டு மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்
8) N-gramms-க்கு மெருகிடுதல் (Soothing for N-grams)

இயல் 5

தமிழ்த் தரவுத்தொகுதியில் தொடர் பகுத்தல்

5.1 முன்னுரை

தொடர் பகுத்தல் (chunking) என்பது வாக்கியங்களைத் தொடர்களாகப் பகுத்தலின் முக்கியக் கூறாகும். பகுத்தல் என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்க ஆய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும். வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுத்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளில் கொண்டு வருகிறது. வாக்கிய நிலைப் பகுத்தல் சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது

கிளைவங்கி (Tree bank) என்றழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா.

'Pearl sat on a chain' என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – Sat VVI] [PP on II]

[NP a – AJI] [Chair – NNI]

[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.
2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பகுத்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்துக் குறித்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகின்றது. பகுத்துக் குறித்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்துக் குறித்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்தலின் குறைப்பாடு பகுத்தல், தரவுத்தொகுதி திரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள் சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

அடிப்படையில் தொடர் பகுப்பான் (chunker) வாக்கியத்தை மேலுறல் செய்யாத முக்கியமான தொடர்களாகப் பிரித்து அதற்குப் புலக்குறிப்பு (label) தருகிறது. தொடர்பகுப்பான் அவற்றின் துல்லியமான வெளிப்யீடு அடிப்படையிலும் எவ்வாறு தொடர் பகுதிகள் விளக்கப்படுகிறது என்பதன் அடிப்படையிலும் வேறுபடுகின்றன. பல தொடர் பகுப்பான்கள் எளிய தொடர்களைப் பகுத்தலைவிடக் கூடுதலாகச் செய்கிறது. பிற பெயர்த்தொடர்களை மட்டும் கண்டுபிடிக்கிறது. தொடர் பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலுக்கும், முழு பகுத்துக் குறித்தலுக்கும் இடையில் வருகிறது. தனிப்பட்ட தொடர் பகுதிகளின் அமைப்பை விளக்குவது எளிது. ஆனால்

அவற்றிற்கிடையிலான உறவுகளை விளக்குவது கடினம். இது தனிப்பட்ட சொல் வகுப்புகளைச் சார்ந்திருக்கிறது. இவ்வாறு தொடர்களைப் பகுத்தில் தற்பொழுது இருக்கின்ற முறைக்கும் முன்மாதிரி வெளிப்யீட்டு ஆய்வுக்கும் இடையில் உள்ள சலுகையான முறையாகும். தொடர் பகுப்பான்கள் வாக்கியங்களைப் பிரித்து அடையாளப்படுத்துகிறது. பெரும்பாலான தொடர் பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலில் உள்ள தகவலைப் பயன்படுத்துகிறது. பிற உண்மையான சொற்களைப் பார்க்கிறது.

சொல் வகைப்பாட்டு அடையாளப்படுத்தி சொல் மற்றும் சூழலுக்கான தகவலைப் பயன்படுத்தி இம்மயக்கங்களை நீக்குகின்றது. சொல் வகைப்பாட்டுத் தகவல் பகுத்துக்குறித்த அமைப்பில் உறுப்பின் நிரல் அல்லது இடத்தைக் கண்டுகொள்ள தேவைப்படுகிறது. பண்புக்கூறு தகவல்கள் நிரப்பியையும் உடன்பாட்டையும் பரிசோதிக்கப் பயன்படுகிறது.

தமிழில் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமானது. தொடரமைப்பு விதிகளின் எல்லா விதமான வரிசை மாற்றங்களுக்கும் வேண்டி எழுதப்பட வேண்டும் என்று அவசியப்படும் இவ்வகங்களின் காரணமாக இப்பகுத்துக் குறிப்பான்கள் தேவையில்லாமல் விதிகளின் எண்ணிக்கையைப் பெருக்கிக் கொள்ளும். பகுத்துக்குறிப்பான் மீது கூடதலான சுமை புகுத்தப்படும். தற்போதைய ஆய்வில் பகுத்துக்குறித்தலுக்கு முன் வாக்கியத்தை ஒரு குறிப்பிட்ட நிலை பேறாக்கம் செய்யும் நடத்தை பின் பற்றப்படுகிறது.

வினைதான் வாக்கியத்தின் முக்கிய உறுப்பாகும். வினைகள் அவற்றின் பொருண்மை இயல்பு அடிப்படையில் வேறுபட்ட பங்களிப்பாளர் அமைப்பை ஏற்கும். இப்பங்கெடுப்பாளர் அமைப்புகள் பெயர்த்தொடர்களின் வேற்றுமை உருபுகளை விளக்கும். அவை வினைக்கு நேரடியான மற்றும் நேரடியல்லாத செயப்படுபொருளாக அமையும். பகுத்துக்குறிப்பான் வாக்கியத்தில் வினைக்குப்

பொருத்தமான பங்கெடுப்பாளர் அமைப்பை பரிசோதிப்பதைக் கவனிக்க வேண்டும். மேலும், தமிழில் எழுவாயின் பால்-எண்-இட பண்புக்கூறுகள் பயனிலையுடன் இயைபுடையதாக இருக்க வேண்டும். ஒரு பகுத்துக்குறிப்பான் வாக்கியத்தைப் பகுத்துக்குறிக்கும்போது வாக்கியத்தைச் சரியாக மதிப்பிட இவ் இயைபு பண்புக்கூறைப் பரிசோதிக்க வேண்டும்.

5.2 பகுத்தல்

பகுத்தல் (parsing) இயந்திரமொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுத்தல் போன்ற பயன்பாடுகளில் முக்கியமான செயல்பாடாகும். சில பயன்பாடுகள் வாக்கியங்களின் முழுமையற்ற/ஆழமில்லாப் பகுத்தலை (shallow parsing) வேண்டுகிறது. ஆங்கிலம் மற்றும் பிற ஐரொப்பிய மொழி வாக்கியங்களின் பகுத்தலுக்குப் பல முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இந்திய மொழிகளில் இந்தி மற்றும் வங்காள மொழிகளுக்கு பகுப்பான்கள் இருந்தாலும் தமிழைப் பொறுத்தவரையில் கூறத்தக்க பகுப்பான் இல்லை. தமிழில் பல கோட்பாட்டு மொழியியல் அடிப்படையிலான தொடரியல் ஆய்வுகள் செய்யப்பட்டுள்ளன. இருப்பினும், தமிழ் வாக்கியங்களின் பகுத்தலுக்கு வேண்டிய கணினியியல் சார் சட்டகத்தை உருவாக்க அச்சிட்டு வெளிப்யிடப்பட்ட புத்தகங்கள் இல்லை. தற்போதைய ஆய்வு தமிழ் வாக்கியங்களின் பகுத்தலுக்கு கணினி அடிப்படையிலான இலக்கணச் சட்டகம் உருவாக்குவதில் கவனம் செலுத்தப்படுகின்றது.

5.3 ஆழமில்லாப் பகுத்தல்

நாம் தொடரியல் என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளிப்யீட்டை விடக் குறைவான நிறைவுடைய ஆய்வைக் குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகிறோம். ஆய்விலிருந்து கிடைக்கும்

வெளிப்யீடு தொடரமைப்புக் கிளைப்படம் அல்ல. ஒரு பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அவற்றின் அக அமைப்பையும், செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். மற்றொரு வகையான ஆய்வு முதன்மை வினை மற்றும் அவற்றின் நேரடியான பங்கெடுப்பாளர் போன்ற சொற்களின் சில செயல்பாட்டுப் பங்களிப்பை அடையாளம் காணும். பகுப்பாய்வு ஒழுங்குமுறை உருபயனியல் ஆய்வின் மீதும் பொருண்மை மயக்க நீக்கலின் மீதும் இயல்பாக வேலை செய்யும். இதன் அடிப்படை நோக்கம் தலைச்சொல் (Head word) உருபனியல் தகவல் மற்றும் சொல்வரிசை அமைப்பு இவற்றில் இருந்து சாத்தியமான தொடரியல் அமைப்பைப் (Syntantic structure) பெறுவதாகும். முன்மாதிரியாக பகுப்பாய்வுத் தொடர்களையும் தலை-அடை (head-medifier) உறவுகளையும் கண்டுபிடிப்பதை நோக்கமாகக் கொள்கிறது. பெரிய உரை தரவுத்தொகுதிகளின் (corpus) பயன்பாட்டில் ஆழமில்லாப் பகுப்பாய்வி பயன்படுத்தப்படுகிறது. பெரும்பாலும் பகுப்பாய்விற்கு எல்லாச் சிக்கல்களுக்கும் தீர்வு காண்பதற்கு போதிய திறன் இல்லாவிட்டால் பகுதியான ஆய்வுகள் அனுமதிக்கப்படும். சர்ச் (Church.1988) பொருத்தமான அடைப்புக் குறிகளைச் செருகி எளிய பெயர்த் தொடர்களை அடையாளம் காணும் புள்ளியல் (Stochastic) வழியமைப்பை வடிவமைத்துள்ளார். ஒப்னெய் (Obney, 1991) உளமொழியியல் சான்றுகள் அடிப்படையிலும், நடைமுறை பயன்பாடுகள் அடிப்படையிலும் ஆழமில்லாப் பகுப்பாய்வின் முக்கியத்துவத்தை முதலில் விவாதித்தவர் என்று மதிக்கத்தக்கவர் ஆவார். அவர் தமது அணுகுமுறையில் ஆழமில்லாப் பகுப்பாய்வைப் பெறுவதற்காக முற்றுநிலை மாறிகளைப் பயன்படுத்தினார். ஆழமில்லாப் பகுப்பாய்வி சொல்வகைப்பாடு அடையாளப்படுத்தல், தொடர் உறுப்பு வரையறுத்தல், தொடர் உறவுகளைக் கண்டுபிடித்தல் என்ற முன்மாதிரி தொகுப்புகளை உள்ளடக்கியது.

5.3.1 சொல்வகை அடையாளப்படுத்தல்

ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் சொல்வகை அடையாளப்படுத்தி அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை) உறுதிசெய்யும் சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தல் (Parts of Speech tagging) இயற்கைமொழி ஆய்வில் மிக அறியப்பட்ட சிக்கலாகும். இதில் இயந்திர கற்றல் அணுகுமுறைகள் (machine language approaches) பொதுவாகப் பயன்படுத்தப்படுகிறது.

5.3.2 தொடர் பகுத்தல்

சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்புகளும் தரப்படுகையில் எந்தச் சொற்கள் தொடர்களாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பியத்தொடர் போன்றவைகளாக) குழுமப்பட வேண்டும் என்று தொடர் பகுத்தல் (Chunking) செயல்பாட்டால் தீர்மானிக்கப்படுகிறது.

தொடர் பகுத்தலுடன் பின்வரும் கேள்வி தொடர்புள்ளது. தொடர் எதைக் கொண்டிருக்கும்? ஒரு மாதிரி தொடர் ஒரு தனி நிலைப் பொருளடக்கச் சொல்லைச் சுற்றி ஒரு கூட்டம் செயல் பாட்டுச் சொற்களைக் கொண்டிருக்கும் (Abney, 1991). தொடர்கள் இயல்பாக ஒன்றுக்கொன்று இயைபுள்ள சொற்களின் குழுமமாக எடுத்துக்கொள்ளப்படுகின்றது.

அடுத்த கேள்வி 'நம்முடைய தேவைக்காக இந்த இயைபுள்ள சொற்களின் குழுவின் எல்லைகளை எவ்வாறு வரையறை விளக்கம் செய்யவியலும்?' என்பதாகும். எடுத்துக்காட்டாக, பின்வரும் இணைகளில் எது சரியான தொடர்? என்ற ஐயம்.

1. ((மேசையின் மேல்)) அல்லது
2. ((மேசையின்)) ((மேல்))

5.3.3 தொடர் உறவைக் கண்டுபிடித்தல்

ஒரு வாக்கியத்தில் உள்ள தொடர்கள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனிலை, இடம் போன்றவை) தீர்மானிக்கப்பட வேண்டும் என்ற நிலையில் உறவைக் கண்டுபிடித்தல் (Relation finding) முக்கியமான செயல்பாடாகும்.

ஆழமில்லாப் பகுப்பாய்விகள் இயற்கை மொழி முழுவதையும் கையாள வேண்டி இருப்பதால் அவை பெரிதாகவும் பெரும்பாலும் ஆயிரக்கணக்கான விதிகளையும் கொண்டிருக்கின்றன. இவ்விதிகளில் பல விதிவிலக்கானவை என்பதன் அடிப்படையில் இவை வலிமை குறைந்தவையாக இருக்கின்றன. ஆழமில்லாப் பகுப்பாய்விகள் இயந்திர கற்றல் அடிப்படையிலான தொழில்நுட்பங்களைப் (techniques) பயன்படுத்தித் தானியக்கமாக உருவாக்கப் படுகின்றன.

5.4 பெயர்த்தொடர் பகுத்தல்

பெயர்த்தொடர் பகுத்தல் வாக்கியங்களிலிருந்து பெயர்த் தொடர்களைப் பிரித்தெடுப்பதைக் குறிப்பிடுகின்றது. பொதுவான பகுத்தலை விடப் பெயர்த்தொடர் பகுத்தல் எளிதாக இருந்தாலும் ஒரு துல்லியமான மற்றும் திறமைவாய்ந்த பெயர்த்தொடர் பகுத்தல் பல பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது என்ற உண்மையால் மிக முக்கியத்துவம் வாய்ந்ததாகக் கருதப்படுகின்றது.

பெயர்த்தொடர்களைப் பனுவல்களைப் பகுப்பதற்கு முன் முன்பகுப்பாய்வுக் கருவியாகப் (pre-processing tool) பயன்படுத்த இயலும். இயற்கை மொழியின் கூடுதலான பொருண்மை மயக்கம் காரணமாக பனுவலின் சரியான பகுத்துக்குறித்தல் மிகக் கலவைத் தன்மையானதாக (சிக்கலானதாக) இருக்கின்றது. இந்நேர்வுகளில் தொடர்களைப் பகுத்தலை இம்மயக்கங்களை ஓரளவு தீர்ப்பதற்கான முன்பகுப்பாய்வுக கருவியாகப் பயன்படுத்த இயலும். இப்பயன்பாட்டில் தொடர் பகுத்தலைச் சொற்களுக்குப் பதிலாகத் தொடர் பகுதிகளைச் சார்ந்திருக்கும் ஆவணங்களிலிருந்து தரவை மீட்பதற்குப் பயன்படுத்த இயலும். குறிப்பாகப் பெயர்களும் பெயர்த்தொடர்களும் தகவல் மீட்புக்கும் பிரித்தெடுத்தலுக்கும் மிகவும் பயனுள்ளதாகும். இயந்திர மொழிபெயர்ப்பு மீதான சமீபகாலச் செயல்பாடுகள் பயனுள்ள மாற்ற அமைப்பொழுங்குகளை (transfer patterns) ஆக்குவதற்கு இரு மொழிகளிலுள்ள பனுவல்களை (இணை தரவுத்தொகுதிகளை) (parallel corpora) பயன்படுத்துகின்றது. பெயர்த்தொடர்கள் இணை தரவுத்தொகுதிகளில் உள்ள பனுவல்களைப் பொருத்தும் பயன்பாடுகளில் பயன்படுகின்றது. இணை தரவுத்தொகுதிகளில் உள்ள வாக்கியங்களைத் தொடர்ப்பகுப்புத் தகவல்களைப் பயன்படுத்தியும் மூல மொழியிலுள்ள தொடர்களை இலக்கு மொழியிலுள்ள தொடர்களுடன் உறவுபடுத்தியும் பொருத்த வரைவு செய்யலாம். மேலும் பகுக்கப்பட்ட பெயர்த் தொடர்களைத் தரவின் தரமான பகுத்தல் தேவைப்படாத பிற பயன்பாடுகளிலும் பயன்படுத்த இயலும்.

5.5. தமிழ்த் தரவுத் தொகுதியைத் தொடர் பகுத்தல்

5.5.1 தமிழ்த் தொடரியல் அமைப்பு

தமிழைப் பொறுத்தவரையில் அது ஒரு ஒட்டுக்கள் நிறைந்த மொழியாகும். இவ்வொட்டுக்கள் வேர்ச்சொற்களில் சேர்க்கப்பட்டு வேறுபட்ட சொல் வடிவுகளை

உருவாக்கும். ஏனவே தமழுக்கு ஒரு உருபனியல் ஆய்வியை உருவாக்குவது ஒரு சிக்கலான செயல்பாடாகும். பல வாக்கியங்களுக்கு உருபனியல் பகுப்பாய்வியைப் பயன்படுத்தி சொற்களின் அல்லது ஒரு சொல்லின் சொல் வகைப்பாட்டையும் (Parts of speech (POS)) பண்புக்கூறு தகவலையும் அடையாளம் காண இயலும். இருப்பினும், சில சொற்களுக்கு மயக்கமுள்ள சொல்வகைப்பாடும் இலக்கண பண்புக்கூறுகளும் இருக்கும்.

தமிழ்த் தொடரியல் பகுப்பானுக்கு அடிப்படையாக அமையும் தமிழ்த் தொடரியல் குறித்து இங்கு விளக்கப்படும். தமிழ் வாக்கியங்களின் வகைகள் பற்றியும், அவற்றின் அமைப்பு குறித்தும் விளக்கப்படும்.

தமிழ்மொழி வேறுபட்ட சொல் வரிசை அமைப்பைக் கொண்ட தொடர்களைக் கொண்டு வாக்கியங்களை அமைக்கத்தக்க வகையில் அமைந்த மொழியாகும். சொற்களுக்கு இடையிலான தொடர்பினைப் பயன்படுத்தி வாக்கியங்களை இலக்கணரீதியாக அமைக்கிறோம். இலக்கணங்கள் வேறுபட்ட வார்த்தைகளுக்கு இடையே தொடர்பினை உருவாக்கிப் பொருள்தரும் வாக்கியங்களை உருவாக்க உதவுகிறது. இவ்வாறு சொற்களைக் கொண்டு தொடர்களையும் தொடர்களைக் கொண்டு வாக்கியங்களையும் உருவாக்கும் விதிகளைக் கூறுவது தொடரியலாகும்.

ஒன்றுக்கு மேற்பட்ட சொற்கள் அத்தொடர்கள் இணைந்து உருவாகும் வாக்கியங்கள், அவ்வாக்கியங்கள் இணைந்து உருவாகும் கலப்பு, கூட்டு வாக்கியங்கள் ஆகியவற்றின் கட்டமைப்பையும் அவை தரும் பொருண்மையையும் ஆராய்வது தொடரியல் எனப்படும். பொதுவாகத் தொடர்களை ஆராய்வது தொடரியல். அது தொடராக இருக்கலாம் அல்லது வாக்கியங்களாக இருக்கலாம். வாக்கியம் என்பதும் ஒருவகைத் தொடரே.

5.5.1.1 தொடர்கள்

தமிழ் வாக்கியங்களைப் தலைமையாய் வரும் சொல்வகைப்பாடுகளின் அடிப்படையில் பெயர்தொடர், வினைதொடர், பெயரடைத் தொடர், வினையடைத் தொடர், பின்னுருபுத்தொடர் என வகைப்படுத்தலாம்.

5.5.1.1.1 பெயர்த்தொடர்

பெயர்த்தொடர் என்பது ஒரு பெயர்ச்சொல்லைத் தலைமைப் பெயராகக் கொண்ட தொடர் ஆகும். தலைமைச்சொல் பெயராகவோ ஆக்கப்பெயராகவோ (எகா.மயக்கம்) தொகையாகவோ (எகா. பணப்பெட்டி) வினைப்பெயராகவோ, மாற்றுப்பெயராகவோ (எகா.அவன், நான்,நீ போன்று) தொழில் பெயராகவோ (வந்தது, வருவது, வருதல், வந்தமை, வந்தவன் போன்று) இருக்கலாம். பெயர்த்தொடரில் பெயருக்கு முன் அடைகொளி அடை, எண்ணடை, பெயரடை, அளவையடை போன்ற அடைகளும் பெயரெச்சத் தொடரும் வரலாம்.

எ.கா.

பையன், அந்தப் பையன், பெரிய பையன், இரண்டாவது பையன், இரண்டு பையன்கள், என்னுடைய பையன், நேற்று வந்த பையன்.

5.5.1.1.2 வினைத்தொடர்

ஒரு வினைச்சொல்லைத் தலைமைச் சொல்லாகக் கொண்டு உருவாகும் தொடர் வினைத்தொடர் எனப்படும். வினைத்தொடரில் வினைக்கு முன்னர் வேற்றுமைத்தொடர், பின்னுருபுத்தொடர், வினையடை, வினையெச்சத்தொடர் என்பன வரலாம்.

எ.கா.

விழுந்தேன், செய்தேன், அங்கு வந்தேன்

நன்றாக விளையாடினேன், வந்து பார்த்தேன்

போன்றவை வினைத்தொடர்களாகும்.

5.5.1.2 வாக்கிய வகைகள்

வாக்கியங்களைப் பல்வேறு அடிப்படையில் பல வகைகளாகப் பிரிப்பர் இலக்கணயலார். அமைப்பினை அடிப்படையாக வைத்துத் தமிழ் வாக்கியங்களை வினைப்பயனிலை வாக்கியங்கள் என்றும், பெயர்ப் பயனிலை வாக்கியங்கள் என்றும் பிரிப்பர். இதுபோன்று வாக்கியங்களின் சேர்க்கையை அடிப்படையாகக் வைத்து தனி வாக்கியம் அல்லது எளிய வாக்கியம் (Simple sentence) என்றும், கலவை வாக்கியம் (Complex sentence) என்றும் என்றும் பிரிப்பர். வெவ்வேறு வினைகளின் பண்பையும் அதனால் அடையும் மாற்றங்களையும் வைத்து செய்வினை வாக்கியம் (active voice) என்றும், செயப்பாட்டு வினை வாக்கியம் (passive sentence)என்றும் பிரிப்பது உண்டு. இவை மட்டுமின்றி வேறொரு பண்பினை ஒட்டி இலக்கணங்கள் வாக்கியங்களை வினா வாக்கியம் (interrogative sentence) என்றும், செய்தி வாக்கியம் (affirmative sentence) என்றும் பிரிப்பர். இவை மட்டுமல்லாமல் இலக்கணவியலார் ஏவல் வாக்கியம், செய்தி வாக்கியம் என்றும் பிரிப்பர். தமிழில் காணப்படும் வாக்கியங்களை 1. வினைப் பயனிலை வாக்கியம், (கண்ணன் வந்தான்), 2. பெயர்ப் பயனிலை வாக்கியம் (இராமன் ஒரு ஆசிரியர்), 3. பெயரடை பயனிலை வாக்கியம் (இராமன் நல்லவன்), 4. உடைமைப் பயனிலை வாக்கியம் (இராமன் என்னுடையவன், புத்தகம் என்னுடையது) என நான்காகப் பிரித்து இனம் காண்பர் சிலர் (கோதண்டராமன்,1997) .

தொடரியலில் சில முக்கியமான அமைப்புகளை மட்டுமே குறித்துச் சொல்லுகின்ற நிலையில் இதன் கண் ஒருசில அமைப்புகள் மட்டுமே எடுத்துக் கூறப்படுகின்றன.

5.5.1.2..1 அடிப்படை வாக்கிய வகைகள்

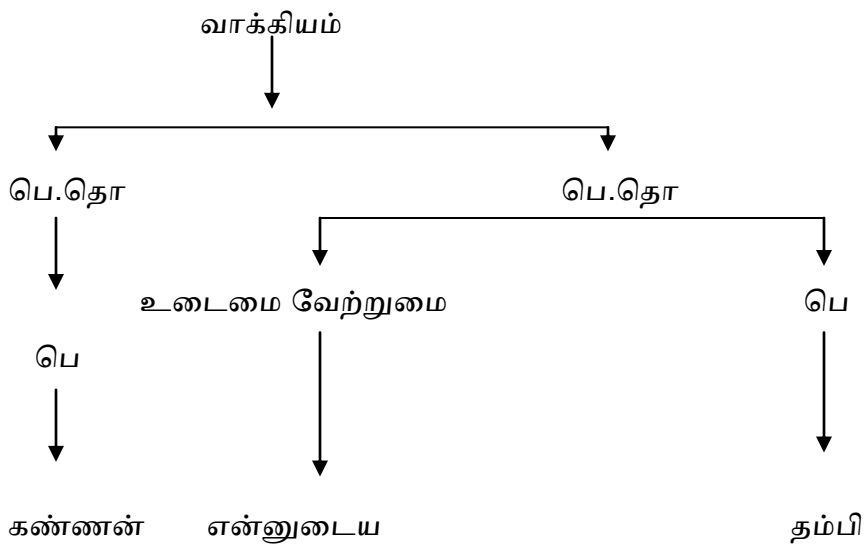
தமிழில் காணப்படும் அடிப்படையான வாக்கியங்கள் 1. பெயர்ப் பயனிலை வாக்கியம் என்றும், 2. வினைப் பயனிலை வாக்கியம் என்றும் இரு வகைகளாகப் பிரிக்கலாம். இவ்விரண்டு வாக்கிய வகைகளிலும் பெயர்த்தொடர் எழுவாய்களாக வரும் இந்நிலையில் இவற்றை

1.பெயர்த்தொடர் + பெயர்த்தொடர் (NP + NP)

2.பெயர்த்தொடர் + வினைத்தொடர் (NP + VP)

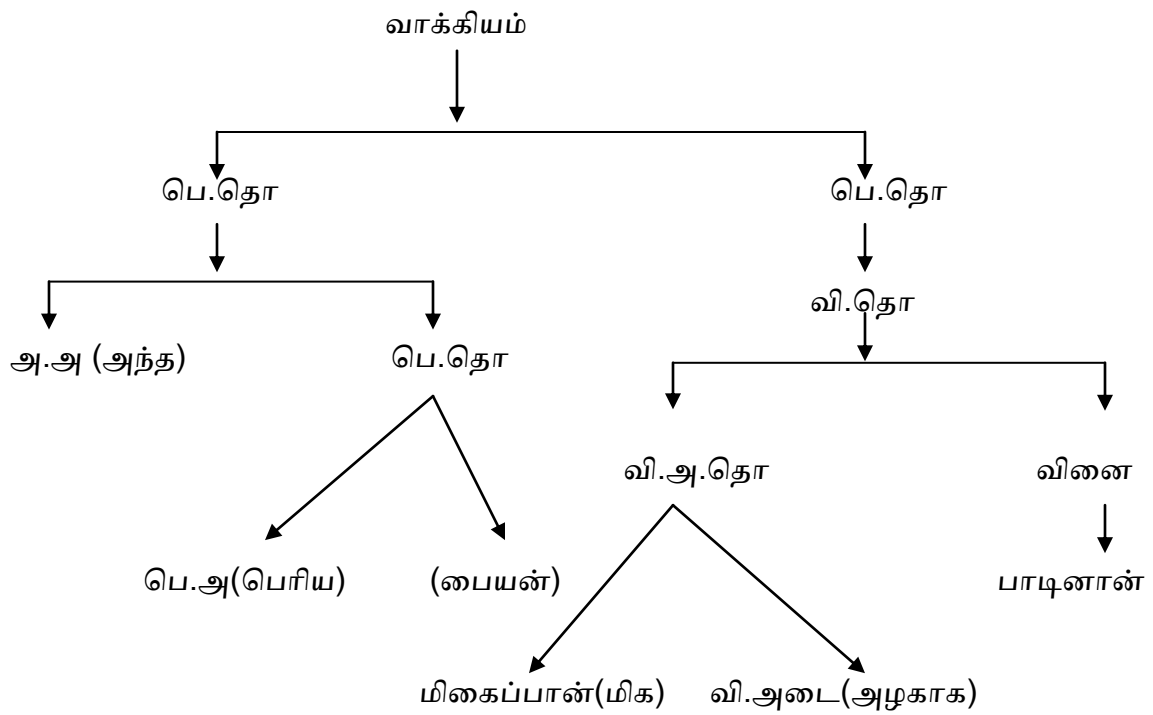
என்ற அமைப்பைக் கொண்டுள்ளவைகளாகப் பிரிக்கலாம்.

ஒரு பெயர்ப் பயனிலைத் தொடர் வாக்கியம் எவ்வாறு உள்ளது என்பது கீழே உள்ள கிளைப்படத்தில் காட்டப்பட்டுள்ளது.



மேலே 'கண்ணன் என்னுடைய தம்பி'என்ற பெயர்ப் பயனிலை வாக்கியத்தின் அமைப்பு காட்டப்பட்டுள்ளது.

வினைப் பயனிலை வாக்கியம் என்பது ஒரு வினையையும் வினையடைத் தொடரையும், மிக என்ற மிகைப்பானையும் கொண்டு வந்துள்ளது.



5.5.1.2..2 வினா வாக்கியங்கள்

வினா வாக்கியம் என்பது தெரியாத ஒன்றை மற்றொருவரிடம் கேட்டுத் தெரிந்து கொள்ளுதற்காகப் பயன்படுத்தப்படும் வாக்கியம் ஆகும். வினாவில் அடிவினா என்றும், ஒட்டுவினா என்றும் இருவகை உள்ளன. அடிவினா யார், என்ன போன்ற வினாச் சொற்களையும் ஒட்டுவினா ஆ என்ற ஒட்டையும் (எ.கா. அவனா? இவனா?) கொண்டிருக்கும்.

வினா வாக்கியங்களும் அவற்றின் விடையாக வரும் செய்தி வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டு விளங்கும்.

ஆம் இராமன் வந்தான்.

அவன் நேற்று வந்தான்.

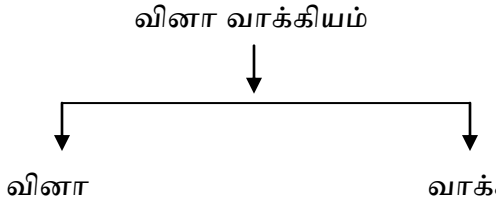
என்ற செய்தி வாக்கியங்களும்,

இராமன் வந்தான். இராமன் வந்தானா?
அவன் நேற்றா வந்தான்?

என்ற ஒட்டு வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டவை. ஒன்றில் வினா ஒட்டு உள்ளது. செய்தி வாக்கியத்தில் அது இல்லை. இதே நிலைதான் அடிவினா வரும் வாக்கியங்களிலும் காணப்படுகின்றது.

இராமன் வந்தான். இராமன் என்று வந்தான்?
இராமன் ஒரு ஆசிரியர். இராமன் யார்?
அவன் நன்கு செய்தான் அவன் எப்படிச் செய்தான்?

போன்ற இணை வாக்கியங்களும் ஒரே மாதிரியான அமைப்பையே கொண்டுள்ளன. ஒன்றில் செய்தி (நேற்று) மற்றொன்றில் வினா (என்று) இதனால் மொழியியலாளர்மள் வினா வாக்கியங்களையும் அதை ஒட்டி வரும் செய்தி வாக்கியங்களையும் ஒரே அமைப்பின் கீழ்க் கொண்டு வந்து வினா வாக்கியத்தைப் பிரித்துக்காட்டும் நிலையில் 'வினா' என்ற முன்னிலையை இணைத்து வேறுபடுத்துவர். இதனைக் கீழ வருமாறு காட்டலாம்.



ஒட்டு வினாவும், அடிவினாவும் ஒன்று வந்த இடத்தில் மற்றொன்று வருவதில்லை. இதனைக் கீழ்வருமாறு காட்டலாம்.

வினா-ஒட்டுவினா,அடிவினா

ஒட்டுவினாப் பொருள்

ஒட்டுவினா ஆம் அல்லது இல்லை என்ற பதில்களை எதிர்பார்ப்பது. முதலில் உள்ளது கேட்போரின் உள்ளத்தில் உள்ளதை உறுதி செய்வது . இரண்டாவது உள்ளதை மறுப்பது.

இராமன் வந்தான்.

இராமன் வந்தானா?

எனத் தமிழில் வருவது காணத்தக்கது. முதல் வினாவின் விடையாக ஆம் (இராமன் வந்தான்) என்றோ இல்லை (இராமன் வரவில்லை) என உறுதி செய்தோ அல்லது மறுத்தோ விடை அமையலாம். இதுபோன்றே அடுத்த வினாவிற்கும் பதில்கள் அமையலாம்.

தமிழில் ஒட்டுவினா, வினா-ஆ என்ற உருபனால் காட்டப்படுகின்றது. பொதுவாக இது வாக்கியங்களில் ஒரே ஒரு இடத்தில் வரும் பல்வேறு இலக்கணக் கூறுகளில் ஏதாவது ஒன்றின் பின்னால் வந்து ஒட்டு வினாப்பொருளைக் காட்டும். இவ்வாறு ஏதாவது ஒன்றின் பின்னால் வருவதை வினாவின் வீச்சு (ளஉடிப்ந) என்பர் மொழியியலாளர்கள்.

இராமன் + வினா + வந்தான் >இராமன் வந்தான்.

இராமன் வந்தான் + ஆ >இராமன் வந்தானா?

என்று வருவது குறிப்பிடத்தக்கது.

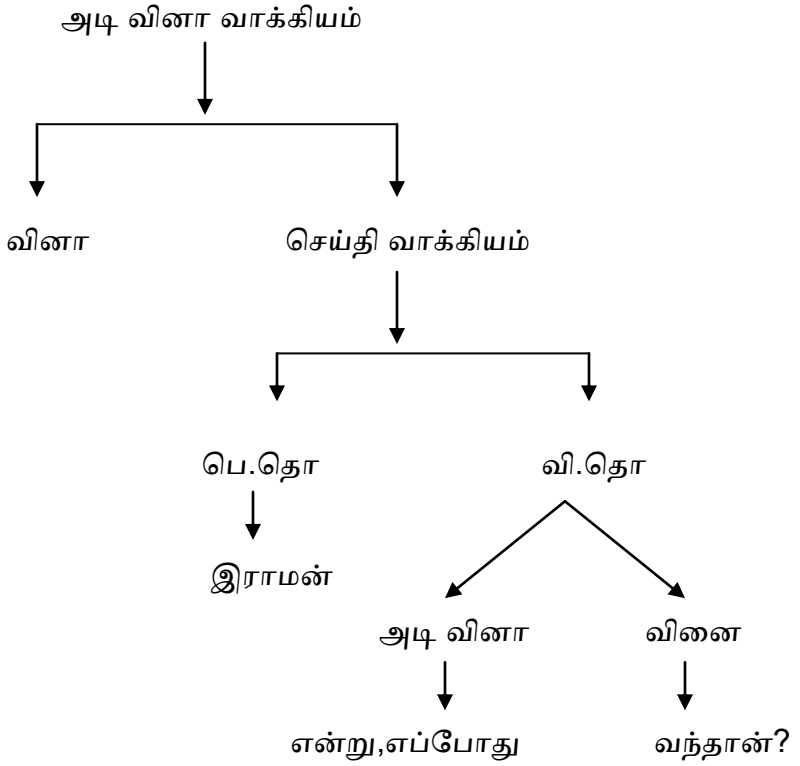
அடி வினா

அடி வினா வாக்கியங்கள் 1.யார், எவர், எவன், என்ன, எது, எவை போன்ற வினாப் பெயர்களையும், 2.எந்த, எத்தனையாவது, எத்தனை, எவ்வளவு, எப்பேர்பட்ட அல்லது எப்படி உள்ள (எத்தகைய, எம்மாதிரி போன்ற) என்ன போன்ற வினாப் பெயரடைகளையும், 3.எத்தனையாவது, எங்கு, எப்படி, எப்போது போன்ற வினா வினையடைகளையும் உள்ளடக்கியவை. இவற்றுள் சில ஒன்றுக்கு மேற்பட்ட நிலைகளிலும் (என்ன, எப்பொழுது) வரும்.

இராமன் நேற்று வந்தான்.

இராமன் என்று வந்தான்?

போன்ற வாக்கியங்கள் ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் 'நேற்று' 'என்று' போன்றவை ஒரே அமைப்பு கொண்ட வாக்கியங்களில் ஒரே இடத்தில் வருபவை ஆகும். எனவே அடி வினா வாக்கியங்களின் அமைப்பைக் கீழ்வருமாறு காட்டலாம்.



என்ற ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் அடி வினா என்ற புனை நிலையில் (hypothetical) கொள்ளப்பட்டது கெட அடி வினா வாக்கியம் உருவாக்கப்பெறுகிறது.

5.5.1.2.3 செய்வினை, செயப்பாட்டு வினை வாக்கியங்கள்

வினாப் பாகுப்பாடு பற்றிக் கூறும்போது வினைகள் செய்வினை என்றும் செயப்பாட்டு வினை என்றும் பகுக்கப்படும் என்று கூறப்பட்டது. செய்வினைப்யன் செயவென் வாய்பாட்டு வினஇயெச்சத்தின் பின்னர் படு அல்லது பெறு என்ற துணை வினைகளை இணைக்கும்போது உண்டாகும் வினைதான் செயப்பாட்டு வினை எனவும் கூறப்படுகிறது. அவ்வாறு உள்ள செயப்பாட்டு வினையைப் பயனிலையாகக் கொண்ட வாக்கியங்கள் செயப்பாட்டுவினை வாக்கியங்கள் ஆகும்.

செயப்படுபொருளை எழுவாயாக ஆக்கியும் செய்பவனை அல்லது மருத்தாவை ஆல் வேற்றுமையாக ஆக்கியும் கூறப்படுவது செயப்பாட்டு வினை வாக்கியம் (Passive Sentence).

இராஜா இராமனை அடித்தான்

என்பது செய்வினை வாக்கியம். இங்கு செய்பவன் எழுவாயாகவும், செய்யப்படுபவன் ஐ வேற்றுமையாகவும் உள்ளது. இதனை உரிய இலக்கண மாற்றத்தால் மாற்றி

இராமன் இராஜாவால் அடிக்கப்பட்டான்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது. இங்குச் செய்வினை வாக்கியம் அடிப்படை வாக்கியமாகவும் (புகைநிலை), செயப்பாட்டு வினை வாக்கியம் வருவிக்கப்பட்ட வாக்கியமாகவும் (புறநிலை) காணப்படுகின்றன. இவற்றிற்கு இடையே காணப்படும் மாற்றத்தைக் கீழ்வருமாறு காட்டலாம்.

பெ.தொ.1+பெ.தொ.2_ஐ+வினை+காலம்+விகுதிகள்

செ.பா

பெ.தொ.2+பெ.தொ.1-ஆல்

வினை. செய.எ.எச்சம்+பகு+காலம்+விகுதிகள்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது.

5.5.1.2.4 கலப்பு வாக்கியம்

ஒரு வாக்கியத்தின் உள்ளே இன்னொரு வாக்கியத்தை முன்னதன் பகுதியாக அல்லது உறுப்பு வாக்கியமாக இணைக்கும்போது உருவாகும் வாக்கியம் கலப்பு வாக்கியம் எனப்படும். எந்த வாக்கியம் தலைமை வாக்கியம் (Matrix sentence) என்றும், இணைக்கப்படுகின்ற வாக்கியம் உறுப்பு வாக்கியம் (Constituent sentence) என்றும் அழைக்கப்படும். எனவே கலப்பு வாக்கியம் என்பது ஒரு தலைமை வாக்கியத்தையும் ஒன்றோ அல்லது ஒன்றுக்கு மேற்பட்ட உறுப்பு வாக்கியத்தையும் கொண்டிருக்கும் வாக்கியம் ஆகும்.

நான் இராமன் வந்தபோது போனேன்

நான் அங்கு வந்த பையனைப் பார்த்தேன்

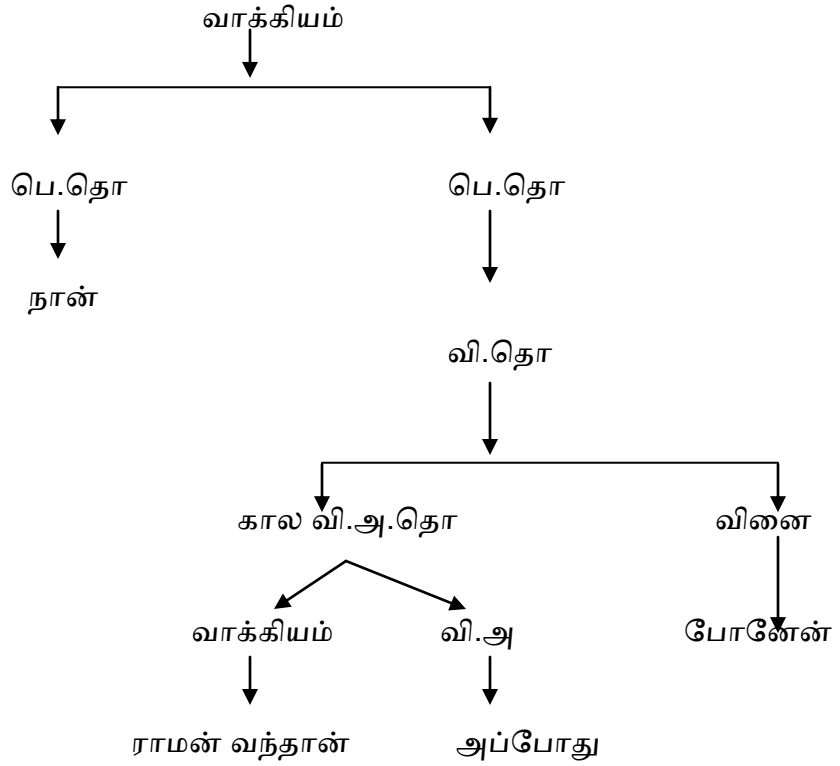
போன்றவை கலப்பு வாக்கியங்கள்.

கால வினையெச்சத் தொடர் வாக்கியம்

இராமன் வந்தபோது நான் போனேன் என்பது ஒரு கலப்பு வாக்கியம். இவ்வாக்கியம் *இராமன் வந்தான், அப்போது நான் போனேன்* என்ற இரு வாக்கியங்கள் ஒன்றில் மற்றொன்று உறுப்பு வாக்கியமாக இணைவதால் ஏற்படும் ஒரு வாக்கியம் ஆகும். இதில் இரண்டாவது வாக்கியம் முக்கிய அல்லது தலைமை வாக்கியமாகவும் முதல் வாக்கியம் அதில் கலந்த உறுப்பு வாக்கியமாகவும் உள்ளன. அப்போது என்ற வினையடையை விளக்கும் நிலையில் அமைந்துள்ளது இராமன் வந்தான் என்ற வாக்கியம்.

புதைநிலை (Deep structure)

இவ்வாக்கியத்தின் புதைநிலையைக் கீழ்வருமாறு காட்டலாம்.



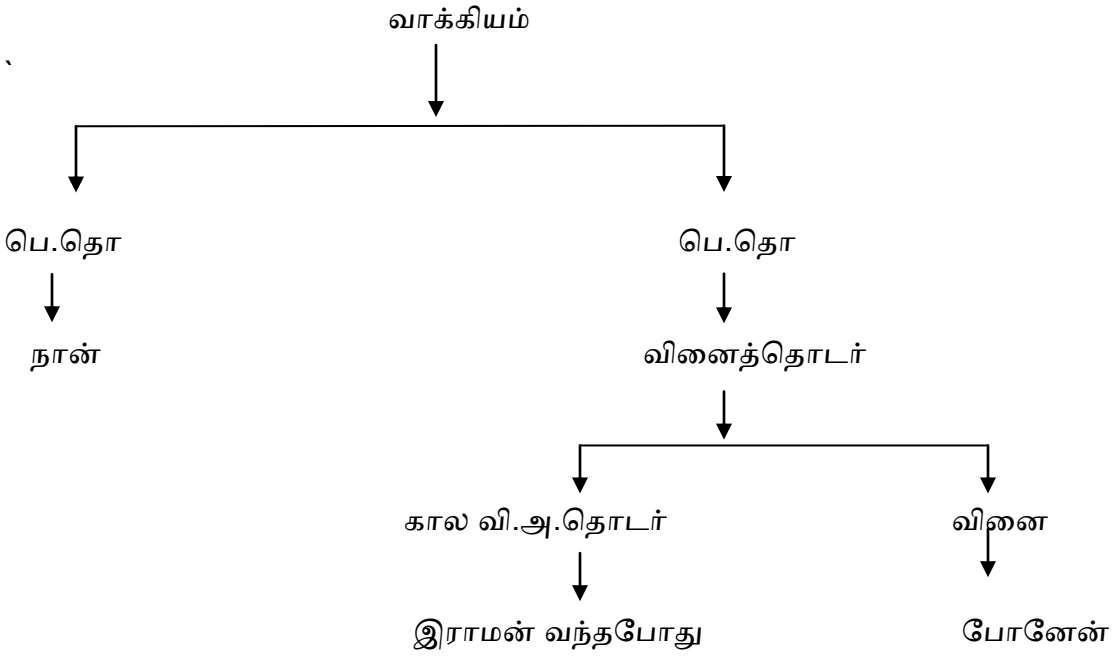
பல்வேறு மாற்றங்கள் (Transformations)

இப்புதைநிலையில் பல்வேறு மாற்றங்கள் நிகழ்கின்றன.

1. வந்தான் என்பது விகுதி. இதுவே பெயரெச்ச விகுதியாகிறது. இதனைப் பெயரெச்ச மாற்றம் (relativisation) எனக் கூறலாம்.
2. அப்போது என்பதில் உள்ள அ என்பது கெடுகின்றது. இதனை அகரச் சுட்டு கெடுதல் எனக் கூறலாம்.

புறநிலை (surface structure)

மேலே காட்டிய மாற்றங்கள் நிகழும்போது நான் இராமன் வந்தபோது போனேன் என்ற புறநிலை வாக்கியம் உருவாகின்றது.



என இதனைக் காட்டலாம்.

பிற கலப்பு வாக்கியங்கள்

பலவிதமான கால எச்சத் தொடர்களும், வேற்றுமைத் தொடர்களும், பின்னுருபுத் தொடர்களும், வினையெச்சத் தொடர்களும், பெயரெச்சத் தொடர்களும் தலைமை வாக்கியத்தின் உறுப்புகளாக வந்து கலப்பு வாக்கியங்களை உருவாக்கும்.

உடனிகழ்ச்சி கலப்பு வாக்கியம்

மணி பாட நான் ஆடினேன்

என்ற வாக்கியத்தில் இரு நிகழ்வுகளைக் குறிக்கும் வாக்கியங்கள் ஒன்றிற்குள் ஒன்றாக உள்ளடக்கப்பட்டு கலப்பு வாக்கியமாக வெளிப்படுகின்றன.

வேற்றுமைத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வேற்றுமைத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

நான் கண்ணன் வந்ததைக் கண்டேன்

பின்னுருபுத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் பின்னுருபுத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நான் வருவதற்கு முன்னர் அவள் வந்து விட்டாள்.

வினையெச்சத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வினையெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

அவள் சாப்பிட்டுவிட்டு தூங்கினாள்

பெயரெச்சத் தொடர் வாக்கியம்

இதில் ஒரு வாக்கியம் பெயரெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நேற்று ஊருக்குப்போன கண்ணன் இன்று பள்ளிக்கு வரவில்லை.

5.5.1.2.5 கூட்டு வாக்கியம்

இரண்டு வாக்கியங்கள் தலைமை வாக்கியம், உறுப்பி வாக்கியம் என்றில்லாமல் ஒத்த நிலையில் இணையும்போது உருவாகும் வாக்கியம் கூட்டு வாக்கியம் எனப்படும்.

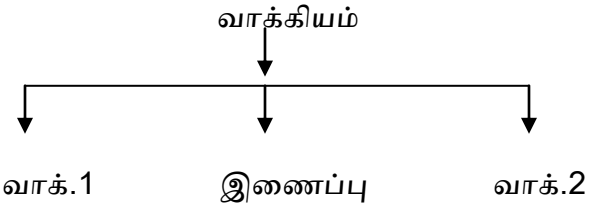
கண்ணன் வந்தான்.

இராமன் வந்தான்.

என்ற இரு வாக்கியங்களும் ஒத்தநிலையில் இணையும்போது

கண்ணனும் இராமனும் வந்தார்கள்.

என்ற கூட்டு வாக்கியம் உருவாகின்றது. இதில் *கண்ணனும் இராமனும்* என்ற பெயர்த்தொடர் எழுவாயாகவும், *வந்தார்கள்* என்பது பயனிலைத் தொடராகவும் உள்ளன. இரண்டு வாக்கியங்களும் கூட்டு வாக்கியத்தில் ஒத்த நிலையில் காணப்படுகின்றன. ஒன்றின் உறுப்பாக மற்றொன்று இல்லை. இத்தகைய இணைப்பு அமைப்பைக் கீழ்க்கண்டவாறு காட்டலாம்.



கூட்டு வாக்கியங்களின் வகைப்பாடு

இரு வாக்கியங்கள் இணைந்து உருவாகும் கூட்டு வாக்கியங்கள் (1) முழுநிலை இணைப்பு வாக்கியங்கள் (conjoined sentences) என்றும், (2) பிரிநிலை இணைப்பு வாக்கியங்கள் (disjunctive) என்றும் இரண்டு வகைப்படும்.

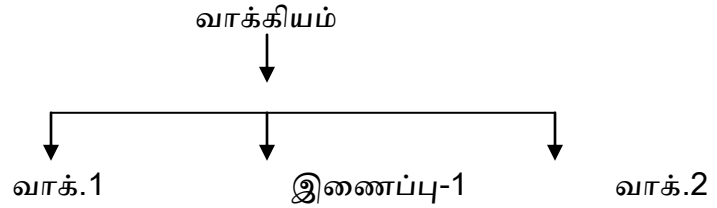
முழுநிலை இணைப்பு

முழுநிலை இணைப்பு என்பது ஒன்றுக்கு மேற்பட்ட வாக்கியங்கள் முழு நிலையில் இணைக்கப்படுவது. அதாவது அதுவா இதுவா என்ற நிலையில்

இணையாது அதுவும், இதுவும் என்ற கூட்டுநிலையில் இணைக்கப்படுவது (conjunctive).

கண்ணனும் இராமனும் வந்தார்கள்.

இதனைப் பின்வருமாறு காட்டலாம்.



பிரிநிலை இணைப்பு வாக்கியம்

இரு வாக்கியங்களை வேறுபடுத்தி அல்லது பிரித்து இணைக்கும் இணைப்பு பிரிநிலை இணைப்பு (disjunctive) எனப்படும். அதாவது 'அதுவா இல்லை இதுவா' என்ற நிலையில் இணைக்கப்படுவது

கண்ணன் வருவான் அல்லது இராமன் வருவான்.
கண்ணன் அல்லது இராமன் வருவான்.
கண்ணனோ அல்லது இராமனோ வருவார்கள்

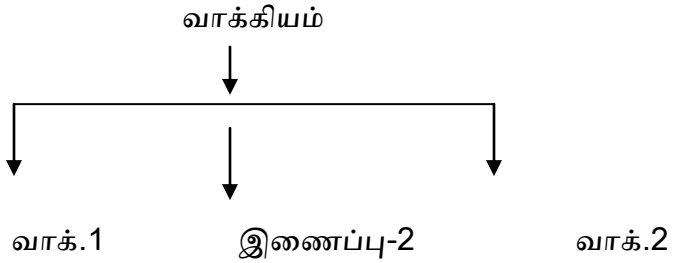
போன்ற வாக்கியங்கள் இப்பிரிவைச் சாரும்.

கண்ணன் வருவான் + இணைப்பு2 + இராமன் வருவான்

என்ற இரு வாக்கியங்களிலிருந்து உருவாவதுதான்.

கண்ணன் வருவான் அல்லது இராமன் வருவான்

போன்ற வாக்கியங்கள் கீழ்வருமாறு காட்டப்படுகிறது.



தமிழ்த்தொடரியல் அமைப்பில் ஒரு பெயர்த் தொடரும் ஒரு வினைத் தொடரும் இணைந்தோ இரு பெயர்த் தொடர்கள் இணைந்தோ ஒரு வாக்கியத்தை உருவாக்கும் வாக்கியங்கள் அவற்றின் அமைப்பின் அடிப்படையில் தனிநிலை வாக்கியம், கூட்டு வாக்கியம் மற்றும் கலவை வாக்கியம் என வேறுபடும். தனிநிலை வாக்கியத்தில் ஒரு வினைத்தொடர் காணப்படும். கூட்டு வாக்கியத்தில் ஒன்றிற்கும் மேற்பட்ட வினைத்தொடர்களோ, பெயர்த்தொடர்களோ இணைப்பான்களால் இணைக்கப்பட்டிருக்கும். கூட்டுநிலை வாக்கியத்தில் ஒன்றுக்கும் மேற்பட்ட வினைத்தொடர்கள் (அதாவது எச்சத் தொடர்கள்) வரும். பொருண்மை அடிப்படையில் வாக்கியங்களைக் கூற்று வாக்கியங்கள், கட்டளை வாக்கியங்கள், வினா வாக்கியங்கள், செய்வினை வாக்கியங்கள், செயப்பாட்டு வினை வாக்கியங்கள் எனப் பகுக்கலாம். ஒரு வாக்கியம் மற்றொரு வாக்கியத்தில் உட்படையாக வருவதன் காரணமாக வாக்கியங்கள் கலவைத்தன்மை பெறுகின்றன. பெயரெச்சத் தொடர்கள் பெயர்த் தொடருக்குள்ளும் வினையெச்சத் தொடர்கள் வினைத் தொடர்க்குள்ளும் உட்படையாக வரும்.

5.5.2 முந்தைய கணினித் தொடரியல் ஆய்வுகள்

முந்தைய பகுத்துக்குறித்தல் அணுகுமுறைகள் பெரும்பாலும் அகராதியால் இயக்கப்பட்டன. இதில் அகராதி உருபனியல் தரவையும், பொருண்மையியல் வர்ணனைகளையும், தொடரியல் அறிவையும் கொண்டிருக்கும். பின்னர் இலக்கண வடிவ வாதங்களின் தொடக்கம் பல வேறுபட்ட கொள்கைகள் அடிப்படையிலான பகுத்துக் குறிப்பான்களைத் தந்தது. காஷ்கெட் (Kashket,1986) என்பவர் வார்ல்பிரி என்ற மொழிக்கு சாம்ஸ்கிலின் (Chomsky) ஆளகை-கட்டுறவுக் கோட்பாடு (government and binding) அடிப்படையில் பகுத்துக்குறிப்பானை உருவாக்கினார். இவ்வணுகுமுறையில் தொடர்கள், தேர்வுக் கொள்கைகள், வேற்றுமை அடையாளப்படுத்துகை, வேற்றுமை தருகை, பங்கெடுப்பாளர் தொடர்பு படுத்துகை என்பனவற்றின் வழி உருவாக்கப்பட்டன. பங்கெடுப்பாளர் தகவல் அகராதியில் சேகரிக்கப்பட்டன. சென்குப்தா (Sengupta, et al., 1997) போன்றோர் கப்லன் மற்றும் பிரஸ்னைன் என்போரின் சொல் செயல்பாட்டு இலக்கணத்தை (Lexical Fuhnctional Grammar)வங்காள மொழிக்குத் தகுந்தவாறு மாற்றி ஒரு பகுத்துக்குறிப்பானை உருவாக்கினார். சொல் செயல்பாட்டு இலக்கணம் ஆங்கிலம் போன்ற இடம்சார் மொழிகளுக்கு (Positional languages) வலுவான வடிவவாதம் ஆகம். இது உறுப்பமைபபையும் இலக்கண உறவுகளையும் கண்டுகொள்கிறது. செங்குப்தாவின் பகுத்துக் குறிப்பான் செய்வினைப்பாடாக உள்ள எளிய வாக்கியங்களுக்கு நல்ல விளவைத் தந்தது. வினிவார்ட்டர் Winiwarter,1995)ஒன்றிணைப்பு இலக்கணத்தை (unofication grammar) விரிவாக்கி ஜோமன் மொழிக்குப் பயன்படுத்தினார். சினிவார்ட்டரின் ஒழுங்கமைப்பில் இலக்கண விதிகள் அருவமாக்கலின் பொருத்தமான மட்டத்தில் இலக்கண விதிக்ள அதராதிக்கு பங்கெடுப்பாளராகச் சேர்ககப்ப்பட்டன. பெக்கா போன்றோர் (Becker et.al, 1994) V_TAG என்ற கிளை இணைப்பு இலக்கணத்திலிருந்து (TREE adjoining) மாறுபட்ட ஒன்றான பலபெயர் கால பகுத்துக்குறித்தல் வழிமுறை வரைவை உருவாக்கினார் இப்பகுத்துக்குறிப்பானை கிளை இணைப்பு இலக்கண

வடிவவாதம் முற்றிலும் உருவாக்கப்பட்ட மொழிகளுக்கு மட்டுமே பயன்படுத்த இயலும்.

பல பகுப்பான்கள் சார்புக் கொள்கைகள் (dependency principles) அடிப்படையில் உருவாக்ககப்பட்டன. கவிங்டன் (Covington, 2003) உரு சுற்று வரைபடத்தால் இயக்கப்பட்ட சொற்களுக்கு இடையே உள்ள சார்புகளை உருவாக்கி வாக்கியங்களுக்கு அமைப்பைத் தரும் ஒரு பகுப்பானை உருவாக்கினார். இதில் ஒவ்வொரு வில்லும் சொற்களின் சார்புநிலை உறவை உருப்படுத்தம் செய்யும். காரக உறவுகள் வேற்றுமைக் குறிகளுடன் பொருத்தப்படும் பாணினி இலக்கண அடிப்படையில் பாரதி போன்றோர் Gharti, et, al,1993, 2003) கட்டுப்பாடு அடிப்படையிலான பகுப்பான்களின் ஒரு குழுமத்தை உருவாக்கினார். பாரதி ஒழுங்குமுறை பகுக்கும் சிக்கலை முழுஎண் வழியமைப்புச் சிக்கலாக மொழி பெயர்த்துத் தீர்வு காண்கிறது. மூர்த்தி (Murthy,1997) இந்திய மொழிகளுக்கு உரக எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal Clause Structure Grammar) என்ற இலக்கண வடிவ வாதத்தை உருவாக்கினார். இவ்வடிவவாதத்தில் மூன்று மட்ட பகுப்பாய்வு பயன்படுத்தப்படுகிறது. தொடர் அடையாளம் காணல், எச்சத்தொடர் அமைப்பு நிர்ணயம் மற்றும் பங்களிப்பு தருதல். ஒவ்வொரு மட்டத்திற்கும் அதன் பகுப்பாய்வுக்குத் தொடர்புள்ள விதிகள் இருக்கின்றன. உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தில் இத்தொகுதியாக்கம் (modularization) இதைப் பிற மொழிகளுக்கு எளிதாகப் பயன்படுத்த செய்தது. சொல்நிரல் சிக்கலைக் கையாளுவதற்கான ஒரு தீர்மாங்களில் ஒன்று வரிசை மாற்று இலக்கணமாகும் (Permutational Grammar (PG)). (Geg. lkifssib, et, al., 2001) வரிசை மாற்று இலக்கணம் அடிப்படை தொடரமைப்பு விதிகளை அவற்றின் செயல்பாட்டு உருப்படுத்தங்களுடன் குறிப்பிட்டு பின்னர் ஒரே பொருண்மையைத் தரும் பிற எல்லாத் தொடர்ச்சிகளையும் பெறவேண்டி உறுப்புகளை வரிசைமாற்றம் செய்யும். குமாரசண்முகம் மேலே குறிப்பிட்ட வேறுபட்ட அணுகுமுறைகளின் ஒரு

ஒன்றிணைப்பைப் பயன்படுத்துகின்றார். அவர் தொடர்களையும் வாக்கியங்களையும் எச்சத்தொடர்களையும் கண்டுபிடிக்க மூர்த்தியின் உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தைப் போன்று நுட்பங்களைப் (Techniques) பயனபடுத்துகின்றார். சுதந்திர சொல்நிரல் சிக்கலைக் கையாள வரிசைமாற்று இலக்கணத்தைப் போன்று உறுப்புகளை வரிசை மாற்றுவதில்லை. அதற்குப் பதிலாக வாக்கியங்களின் உறுப்புகளை ஒரு குறிப்பிட்ட நிரலில் வரிசைமாற்றம் செய்கிறார். பாரதியின் ஒழுங்கமைப்பில் (Bharati, 2000) உள்ளது போன்று வாக்கியத்தின் எழுவாயையும் செயப்படுபொருளையும் கண்டுபிடிக்க வேற்றுமை உருபுகளைப் பயன்படுத்துகிறார். தமிழின் தொடரியலை விளக்க பொதுவான தொடரமைப்பு இலக்கணததைப் பயன்படுத்துகிறார்.

தமிழ் வாக்கியங்களைப் பகுத்துக் குறிப்பதற்கு வேண்டி ஒரு கணினியியல்சார் சட்டகத்தை உருவாக்கி பயன்படுத்துவதாகும். இதற்கு வேண்டி தமிழ்த்தொடரியலை உருப்படுத்தம் செய்ய தொடரமைப்பு விதிகள் உருவாக்கப்பட்டுள்ளன. நாற்பது மிகக் கூடதலாக நிகழும் வினைகளுக்குப் பங்கெடுப்பாளர் அமைப்பு கண்டுபிடிக்கப் பட்டுள்ளது. பகுத்துக்குறிக்கப்பட்ட வாக்கியங்களின் தொடரியல் அமைப்புகள் பகுக்கப்பட்ட கிளைகளில் உருப்படுத்தம் செய்யப்பட்டுளளன. இவ்வடிவில் அவற்றை இயந்திர மொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுப்பு இவற்றை உள்ளடக்கிய பயன்பாடுகளில் நேரடியாகப் பயன்படுத்த இயலும்.

உள்ளீட்டுத் தமிழ் வாக்கியத்தை முக்கியப் பகுப்பான் பகுப்பதற்கு முன் அது பகுப்பாய்வின் பல மட்டங்களைக் கடந்துசெல்ல வேண்டும். கீழே தந்துள்ள படத்தில் செவ்வகங்கள் பகுப்பாய்வு செய்யும் பகுப்பாய்வுத் தொகுதிகளையும் (analyzer modules) வட்டங்கள் அவை பயன்படுத்தும் பொருத்தமான மூலப்பொருள்களையும் உருப்படுத்தம் செய்யும்.

தமிழுக்கு வேண்டி சில பெயர்த்தொடர் பகுப்பான்கள் உருவாக்கப்பட்டுள்ளன. அவை கீழே விளக்கப்பட்டுள்ளன.

5.5.2.1 AUKBCRC -இன் தமிழுக்கான பெயர்த்தொடர் பகுப்பான்

இவ்வணுகுமுறை விதி அடிப்படையிலானது. இந்நெறிமுறையில் ஒரு தரவுத்தொகுதி எடுக்கப்பட்டு இரண்டோ அதற்கு அதிகமான குமங்களாகப் பிரிக்கப்படும். இப்பிரிக்கப்பட்ட குழுமத்தில் ஒன்று பயிற்சித் தரவாகப் பயன்படுத்தப்படும். பயிற்சித் தரவுக் குழுமம் எடுக்கப்பட்டு பெயர்த்தொடருக்காக மனித முயற்சியால் பகுக்கப்படும். இவ்வாறு ஒரு வாக்கியத்தில் உள்ள பெயர்த் தொடர்களைப் பிரிப்பதற்குப் பயன்படுத்தப்படும் விதிகள் உருவாக்கப்படும். இவ்விதிகள் தொடர் பகுப்பிற்கு அடிப்படையாக அமையும். தொடர் பகுப்பான் வழியமைப்பு இவ்விதிகளைப் பயன்படுத்தி பரிசோதனைத் தரவை பகுக்கும். இவ்விதிகளின் பயன்பாட்டுப் பரப்பு இத்தரவுக்குழுமத்தில் பரிசோதிக்கப்படும். துல்லியமும் மீள அழைத்தலும் இதற்குக் கணிக்கப்படும். விளைவு ஒழுங்குமுறையின் பயன்பாட்டுப் பரப்பை விருத்தி செய்ய கூடதல் விதிகள் தேவையா என்று பரிசோதிக்கப்படும். பின்னர் ஒழுங்குமுறை வேறுபட்ட பிற பயன்பாடுகளுக்குப் பரிசோதிக்கப்படும்.

5.5.2.2 RCILTS –இன் வானவில்

வானவில் இந்திய மொழிகளின் தொழில்நுட்ப முன்னேற்றம் என்ற திட்டத்தின் கீழ் தமிழுக்காகச் செய்யப்பட்ட இயற்கைமொழி ஆய்வுகளில் ஒன்றாகும். இது ஒரு தமிழ் வாக்கியத்தின் தொடரியல் உறுப்புகளை அடையாளம் காண்கிறது. இது பட்டியல் வடிவில் பகுப்புக் கிளைப் படத்தை (parse tree) வெளிப்யிடுகின்றது. இது தனிநிலை வாக்கியங்களைய கூட்டு வாக்கியங்களையும் கையாளுகின்றது. தனிநிலை வாக்கியங்கள் ஒரு வினையையும் பல பெயர்த்

தொடர்களையும், தனிநிலை வினையடைகளையும், பெயரடைகளையும் கொண்டிருக்கும். பல எச்சத் தொடர்கள் உள்ள வாக்கியங்களின் தேர்வுகளில் வானவில் துப்புச்சொற்கள் மற்றும் தொடர்கள் அடிப்படையில் எச்சத் தொடர்களைத் தொடர் குழுக்களாகக் குழுமும். இது தொடரியல் அமைப்பு இலக்கணத்தை உருவாக்கும். வானவில் சொல்நிரலைக் கையாள முன்நோக்குதலைப் பயன்படுத்தும்.

பகுத்தல் என்பது ஒரு வாக்கியத்தை அறிந்துகொண்டு அதற்குத் தொடரியல் அமைப்பைத் தருவதாகும். இது தொடரியல் ஆய்வில் மிகப்பெரிய சிக்கலாகும். இதன் காரணமாக இது இயற்கை மொழி ஆய்வின் ஒரு பெரும்பகுதியாக அமைகிறது.

நிர்ணயிக்கப்பட்ட பகுப்பானின் உள்ளீடு (input) அடையாளப்படுத்தப்பட்ட சொற்களின் நிரலான தொடர்ச்சியும் தொடரமைப்பு விதிகளின் ஒரு குழுமமும் ஆகும். இதன் வெளிப்யீடு (out put) வாக்கியத்தின் பகுக்கப்பட்ட அமைப்பாகும். ஒரு வாக்கியத்தின் பகுக்கப்பட்ட அமைப்பு அதன் எழுவாய், செயப்படுபொருள் மற்றும் பயனிலைகளை அறிந்துகொள்ள உதவுகிறது. மேலும், பகுத்துக் குறிக்கப்பட்ட கிளை அமைப்பு பிற தொடர்களையும் அடை செய்யும் தொடர்களையும் அடையாளம் காணப்படுகிறது.

ஒரு மொழிக்கு ஒரு மாதிரி பகுப்பான் ஒரு அடையாளப்படுத்தியைப் பயன்படுத்தி வாக்கியங்களின் சொல் வகையை அடையாளப் படுத்துகிறது. பின்னர் பகுத்துக்குறித்தும் வழிமுறை வரைவு (algorithm) வாக்கியத்தைப் பகுத்துக்குறித்து அம்மொழி இலக்கணத்திற்குத் தரப்பட்ட தொடரமைப்பு விதிகிளின் அடிப்படையில் பொருத்தமான அமைப்பைத் தரும். பகத்துக்குறித்தலின் போது வழிமுறை வரைவு உறுப்புகளுக்கு இடையேயுள்ள

இலக்கண உடன்பாட்டையும் முக்கிய வினைக்கம் செயப்படுபொருள் தொடர்களுக்கும் இடையே உள்ள சார்பு உறவையும் பரிசோதிக்கும்.

5.5.2.3 குமாரசண்முகத்தின் தொடரியல் பகுப்பாய்வு

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக தமிழில் தொடரியல் பகுப்பாய்வு என்ற ஆய்வேட்டைச் சமர்ப்பித்தார். அதன் சுருக்கம் இங்கு தரப்பட்டுள்ளது.

5.5.2.3.1 வாக்கியங்களின் இலக்கணப் பண்புக்கூறுகள்

வாக்கியத்தில் உள்ள சொற்களின் சொல் வகைப்பாடும் பண்புக்கூறு தகவலும் வாக்கியத்தின் பகுத்துக் குறித்தலுக்கு முக்கியமாகும். பின்வரும் அட்டவணை சொல் வகைப்பாடுகளையும் அவற்றுடன் தொடர்புடைய அடையாளங்களையும் அவற்றின் பண்புக்கூறுகளுடன் காட்டும்.

சொல் வகைப்பாட்டுடன் தொடர்புடைய அடையாளப்படுத்திகள்

வ.எண்	சொல் வகைப்பாடு	அடையாளம்/உருப்படுத்தம்	எடுத்துக்காட்டு
1.	முற்றுவினை	V	பார்த்தான்
2.	பெயர்	N	பையன்
3.	மாற்றுப்பெயர்	P	அவர்
4.	பெயரடை	J	நல்ல
5.	வினையடை	R	வேகமாக
6.	பின்னுருபு	I	முன்னால்
7.	சுட்டடை	D	அந்த

8.	அளவடை	G	பல
9.	பெயர் மாறுபடுத்தி	M	புதுமை
10.	கிழமைப்பெயர்	GN	பொருளின்
11.	முற்றுப்பெறாத வினை	NV	பெற்று
12.	பெயரெச்ச வினை	RV	படித்த
13.	நிரப்பிய வினை	CV	என்று
14.	செயல்பாட்டுச் சொல்	F	வரை

5.5.2.3.2 இலக்கணப் பண்புகளும் அவற்றுடன் தொடர்புடைய அடையாளப்படுத்திகளும்

வ.எண்	பண்புகள்	சொல்வரை/உருப்படுத்தம்
பால்		
1.	ஆண்பால்	+M, -F
2.	பெண்பால்	-M, +F
3.	பலர்பால்	+M, +F
4.	ஒன்றன் பால்	-M, -F
எண்		
5.	ஒருமை	+S
6.	பன்மை	-S
இடம்		

7.	தன்மை	+1, -2
8.	முன்னிலை	-1, +2
9.	படர்க்கை	+1, +2
காலம்		
10.	நிகழ்காலம்	+P, -A
11.	இறந்தகாலம்	-P, +A
12.	எதிர்காலம்	+P, +A
வேற்றுமை		
13.	செயப்படுபொருள் வேற்றுமை	+ACC
14.	கொடை வேற்றுமை	+DAT
15.	இட வேற்றுமை	+LOC
16.	கருவி வேற்றுமை	+INS
17.	உடனிகழ்ச்சி வேற்றுமை	+SOC
18.	நீங்கல் வேற்றுமை	+ABL
19.	பயன் வேற்றுமை	+BEN
20.	கிழமை வேற்றுமை	+GEN

5.4.3.3 உருபனியல் பகுப்பாய்வு

உருபனியல் பகுப்பாய்வு சொற்களைப் பகுக்கும் செயல்பாட்டையும் அவற்றின் ஒட்டு உருபன்பளையும் வேர்களைக் கண்டுபிடிக்கும் செயல்பாட்டையும் உள்ளடக்கும். இவ்வுருபனியல் பகுப்பாய்வினால் ஒரு

சொல்லின் சொல் வகைப்பாட்டையும் அதன் இலக்கணப் பண்புக்கூறுகளையும் அடையாளம் காண இயலும்.

5.4.3.4 சொல் வடிவமும் அதன் அடையாளங்களும்

சொல்வடிவம்	உருபன் விடுவரல்	சொல்லும் அதன் அடையாளப்படுத்தியும்
கேட்டான்	கேள்<verb>+த்த்<past tense>+ஆன்<3rd person, singular,masculine>	கேட்டான் <V:+p,-A, +1.+2 +3, +S, +M, -F>
மரத்தை	மரம் <noun> + ஐ <accusative case>	மரத்தை <N:+ACC, +S, -M, -F>

5.4.3.5 சொல்வகை அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தல் வாக்கியங்களின் உறுப்புகளுக்குச் சரியான அடையாளத்தைத் தரும் செயல்பாடாகும். தமிழ் போன்ற மொழிகளுக்கு உருபனியல் பகுப்பாய்வி 70% துல்லியத்துடன் சொல்வகைப்பாட்டை கண்டுபிடிக்க இயலும். இருப்பினும் பல நிகழ்வுகளில் உருபனியல் பகுப்பாய்வி ஒரு சொல்லுக்கு ஒரு அடையாளத்தை மயக்கமின்றி தர இயலாது. இம்மாதிரியான மயக்கங்களை நீக்கக் குமாரசண்முகம் உருபனியல் பகுப்பாய்விற்குப் பின்னர் சொல்வகை அடையாளப்படுத்தியைப் பயன்படுத்துகிறார். சொல்வகை அடையாளப்படுத்தி உருபனியல் ஆய்வியின் விடுவரல் (output) ஏற்று இயலுகின்ற அளவு சொல்மட்ட விதிகளைப் பயன்படுத்தி மயக்கத்தை நீக்குகிறது. எஞ்சுவது பகுத்துக்குறிப்பானால் கவனித்துக் கொள்ளப்படும்.

5.4.3.6 இயல்பாக்கம் (Normalization)

தமிழ் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமுள்ள மொழியாகையால் இறுதி வினைக்கு முன்னர் வரும் பெ.தொ பங்கெடுப்பாளர்கள் இறுதி வினைக்கு முன் எந்த வரிசையிலும் வந்து ஒரே பொருளைத் தரும்.

எ.கா

ஆசிரியர் குழந்தைக்குப் பரிசைக் கொடுத்தார்.

ஆசிரியர் பரிசைக் குழந்தைக்குக் கொடுத்தார்.

பரிசை ஆசிரியர் குழந்தைக்குக் கொடுத்தார்.

பரிசைக் குழந்தைக்கு ஆசிரியர் கொடுத்தார்.

குழந்தைக்குப் பரிசை ஆசிரியர் கொடுத்தார்.

குழந்தைக்கு ஆசிரியர் பரிசைக் கொடுத்தார்.

குமாரசண்முகம் இயல்பாக்கத்திற்காக வாக்கியங்களில் தொடர்கள் ஒரு அனுமதிக்கப்பட்ட எல்லைக்குள் ஏற்கெனவே தீர்மானிக்கப்பட்ட வரிசைகளில் வருவதாக எடுத்துக் கொள்கிறார். இயல்பாக்கி வாக்கியத்தை வரிசையாக ஆய்ந்து தொடர் எல்லைகளை அடையாளம் காண்கின்றது. அதைத் தொடர்ந்து அது எச்சத் தொடர்களையும், பெயர்த்தொடர்களையும் இறுதி வினைக்கு முன் ஒரு குறிப்பிட்ட நிலையில் வரிசைப்படுத்துகிறது. பெயர்த்தொடர்களின் மற்றும் எச்சத் தொடர்களின் எல்லைகள் கண்டுபிடிக்கப்பட்ட பின் இயல்பாக்கி இவற்றை வலமிருந்து இடமாக பின்வரும் முன்நிர்ணயிக்கப்பட்ட நிரலில் மறுவரிசைப்படுத்தும்.

0. வினை,முற்றுப்பெறா வினை,பெயரெச்ச வினை,நிரப்பு வினை

1. பெ.தொ+செயப்படுபொருள் வேற்றுமை

2. பெ.தொ+கொடை வேற்றுமை

3. பின்னுருபு+இட வேற்றுமை

4. பின்னுருபு+கருவி வேற்றுமை

5. பின்னுருபு+உடனிகழ்ச்சி வேற்றுமை

6. பின்னுருபு+நீங்கல் வேற்றுமை

7. பின்னுருபு+பயன் வேற்றுமை

8. வினையடை+கிழமை வேற்றுமை

9. எழுவாய்

எடுத்துக்காட்டாக முன்னர் பட்டியலிடப்பட்ட வாக்கியங்களில் எது தரப்பட்டாலும் இயல்பாக்கி பின்வரும் வெளிப்யீட்டைத் தரும்.

ஆசிரியர் குழந்தைகளுக்குப் பரிசைக் கொடுத்தார்.

5.4.3.7 இலக்கணச்சட்டகம்

பகுத்துக் குறித்தலுக்குக் குமாரசண்முகம் தொடரமைப்பு இலக்கணத்தைப் பின்பற்றுகிறார். சொல் வகைப்பாடுகளில் இருந்து கிளைகளை உருவாக்க X-பார் கோட்பாட்டை ஒக்கும் அணுகுமுறையைப் பின்பற்றுகிறார்.

தொடரமைப்பு விதிகள் ஒரு குழும விதிகளைக் கொண்டது. சண்முகசுந்தரம் பின்பற்றுகின்ற விதிகள் கீழே பட்டியலிடப்பட்டுள்ளன.

Sl.No	Rules	Examples
	Quantifier Phrase	Q->Quantifier

	Rules	
1	Q '=> Q	Q(Q(சில)) Q(Q(few))
2	QP => Q	QP (Q(சில)) QP (Q(few))
	Adjectival Pharase (ADJP) Rules	j->Adjective
3	J => J	J'(J(நல்ல)) J'(J(good))
4	ADJP => J	ADJP(நல்ல)) ADJP(J'(good))
	Modifier Phrase (MP) Rules	M->Noun modifier
5	M => M	M'(M'(கலை)) M'(M'(art))
6	MP => M	MP'(MP'(கலை)) MP (MP'(art))
	Genitive Phrase (GP) Rules	GN->Genitive noun D->Determiner RC->Relative Clause
7	GN => *GN	GN'(GN'(மாணவனுடைய)) GN'(GN'(student's))
8	N =>	GN'(MP(கல்லூரி) GN'(மாணவனுடைய))

	MP*GN	GN'(MP'(college)GN'(student's))
9	GN => ADJP*GN	GN'(ADJP (நல்ல) GN'(மாணவனுடைய)) GN'(ADJP(good) GN'(student's))
10	GN'=> RC*GN	GN'(RC(<trace>படத்தைப் பார்த்த) GN'(பையனுடைய)) RC(<trace>movie saw+RP)GN'(boy's))
11	GN'=> GP*GN	GN'GP(மாணவனுடைய)GN'(புத்தகத்தின்)) GN'(GP(student's) GN'(book's))
12	GP=>*GN	GN'(GP' மாணவனுடைய)) GP'(GP'(student's))
13	GP=>D*GN'	GP(D(அந்த)GN'(மாணவனுடைய)) GP(D(that)GN'(student's))
	Noun Phrase (NP) Rules	**P-> Pronoun N-> Noun** **F-> Functional word**
14	N' => *N	N'(N(கல்லூரி)) N'(N(college))
15	N' => P *N	N'(P(அவன்) N'(P(he))
16	N' => MP *N'	N'(MP(fiy)N'(கல்லூரி)) N'(MP(art)N'(collegep))
17	N' => QP *N'	N'(QP(சில)N'(கல்லூரி)) N'(QP(few)N'(college))
18	N' => ADJP*N'	N'(ADJP(நல்ல)N'(கல்லூரி)) N'(ADJP(good)N'(college))
19	N' => RC*N'	N'(RC(<tarace> படத்தைப் பார்த்த)N'(பையன்))

		N’(RC(<tarace>movie saw+RP)N’(boy))
20	N’ => GP*N’	N’(GP(மாணவனுடைய)N’(புத்தகம்)) N’(GP(student’s)N’(book))
21	N’ => N’F	N’(N’(வீடு)F(வரை)) N’(N’(house)F(till))
22	NP => *N’	N’(N’(கல்லூரி)) N’(N’(college))
23	NP => D*N’	NP(D(அந்த)N’(கல்லூரி)) NP(D(that)N’(college))
	Postposition Phrase (PP) Rules	**I -> Postposition** **N’ -> N’wutg itger tgan Accusative and Dative cases**
24	PP => *N’	PP(N’(கல்லூரியில்)) PP(N’(college in))
25	PP => D*N’	PP(D(அந்த)N’(கல்லூரியில்)) PP(D(that)N’(college in))
26	P => *N’I	PP(N’(வீட்டிற்கு)I(முன்)) (N’(house in)I(front of))
	Adverbial Pharase (ADVP) Rules	**R -> Adverb**
27	R’ => R	R’(R(மெதுவாக)) R’(R(slowly))
28	R’=> QPR’	R’(QP(மிக)R’(மெதுவாக))

		R’(QP(very)R’(slowly))
29	ADVP => R’	ADVP(R’(மெதுவாக)) ADVP(R’(slowly))
	Verrb Phrase (VP) Rules	**V -> Verb**
30	V’ => *V	V’(V(பார்த்தான்)) V’(V(saw))
31	V’ => NP *V	V’PP(காடுகளில்)V(காணப்படுகிறது)) V’PP(in forest)V(seen))
32	V’ => PP*V	V’(NP(படத்தைப்)V(பார்த்தான்)) V’(NP(movie)V(saw))
33	V’ => PP PP *V	V’(ஜோடியுடன்PP(காடுகளில் V(காணப்படுகிறது)) V’(with pair)PP(in forest)V(seen))
34	V’ => NP NP *V	V’(NP(அவனுக்கு)NP(புத்தகத்தைக்)V(கொடுத்தேன்)) V’(NP(him)NP(book)V(gave))
35	V’ => PP NP *V	V’(PP(திரையரங்கில்)NP(படத்தைப்)V(பார்த்தான்)) V’(PP(in theatre)NP(movie)V(saw))
36	V’ => PP PP NP *V	V’(PP(திரையரங்கில்)PP(நண்பருடன்)NP((படத்தைப்)V(பார்த்தான்)) V’(PP(in theatre)PP(with friend)NP(movie)V(saw))
37	V’ => PP NP NP *V	V’(PP(வீட்டில்)NP(அவனுக்கு)NP(புத்தகத்தைக்)V(கொடுத்தேன்)) V’(PP(at home)NP(him)NP(book)V(gave))
38	V’ => NC *V’	V’(NC(படத்தைப் பார்த்து)V’(கருத்தைச் சொன்னான்)) V’(NC(movie saw+VBP)V’(comments said))
39	V’ => ADVP *V’	V’(ADVP(கவனமாகப்V’(பார்த்தான்)) V’(ADVP(carfully)V’(saw))

40	V' => *V'	VP(V'(பார்த்தான்)) VP(V'(saw))
	Nonfinite Verb Phrase (NVP) Rules	**NV -> Nonfinite Verb NC->Non-finite clause VBP-> Verbal Participial**
41	NV' => *NV	NV'(NV(பார்த்து)) NV'(NV(saw))
42	NV' => PP *NV	NV'(PP(காடுகளில்NV(காணப்பட்டு)) NV'(PP(in forest)NV(seen+VBP))
43	NV' => NP *NV	NV'(NP(படத்தைப்)NV(பார்த்து)) NV'(NP(movie)NV(saw+VBP))
44	NV' => PP PP *NV	NV'(PP(ஜோடியுடன்)PP(காடுகளில்)NV(காணப்பட்டு)) NV'(PP(with pair)PP(in forest)NV(seen+VBP))
45	NV' => NP NP *NV	NV'(NP அவனுக்கு)NP(புத்தகத்தைக்)NV(கொடுத்து)) NV'(NP(him)NP(book)NV(gave+VBP))
46	NV' => PP NP *NV	NV'(PP(திரையரங்கில்)PP(படத்தைப்)NV(பார்த்து)) NV'(PP(in theatre)PP(movie)NV(saw+VBP))
47	NV' => PP PP NP *NV	NV'(PP(திரையரங்கில்)PP(நண்பனுடன்PP(படத்தைப்)NV(பார்த்து)) NV'(PP(in theatre)P P(withfriend)PP(movie)NV(saw+VBP))
48	NV' => PP NP NP *NV	NV'(PP(வீட்டில்NP(அவனுக்கு)NP(புத்தகத்தைக்)NV(கொடுத்து)) NV'(PP(at home)NP(him)NP(book)NV(gave+BVP))
49	NV' => NC *NV'	NV'(NC(படத்தைப் பார்த்து)NV'(கருத்தைச் சொல்லி)) NV'(NC(movie saw+VBP)NV'(comments said+VBP))
50	NV' =>	NV'(ADVP(கவனமாகப்)NV'(பார்த்து))

	ADVP *NV'	NV'(ADVP(carefully)NV'(saw+VBP))
51	NV' => *NV'	NVP(NV'(பார்த்து)) NVP(NV'(saw_VBP))
	Nonfinite Clause (NC) Rules	**CV->Complementizing Verb**
52	NC => *NP *NVP'	NC'(NP(அவன்)NVP(படத்தைப் பார்த்து)) NC'(NP(he)NVP(movie saw+VBP))
53	NVP => *NV'	NC(S(அவன் படத்தைப் பார்த்தான்)CVஎன்று)) NC(S(he movie saw)CV(that))
	Relativizing Verb (RVP) Rules	**RV -> Relativizing Berb** **RP -> Relative Participle**
54	RV' => *RV	RV'(RV(பார்த்த)) RV'(RV(saw+RP))
55	RV' => PP *RV	RV'(PP(காடுகளில்)RV(காணப்பட்ட)) RV'(PP(in forest)RV(seen+RP))
56	RV' => NP *RV	RV'(NP(படத்தைப்RV(பார்த்த)) RV'(NP(movie)RV(saw+RP))
57	RV' => PP PP *RV	RV'(PP(ஜோடியுடன்)PP(காடுகளில்)RV(காணப்பட்ட)) RV'(PP(with pair)PP(in forest)RV(seen+RP))
58	RV' => NP NP *RV	RV'(NP(அவனுக்கு)NP(புத்தகத்தைக்)RV(கொடுத்த)) RV'(NP(him)NP(book)RV(gave+RP))
59	RV' => P NPP *RV	RV'(PP(திரையரங்கில்)PP(படத்தைப்)RV(பார்த்த)) RV'(PP(in theatre)PP(movie)RV(saw+RP))
60	RV' => PP	RV'(PP(திரையரங்கில்)PP(நண்பனுடன்)PP(படத்தைப்)RV(பார்த்த))

	PP NP *RV	RV'(PP(in theatre)P P(withfriend)PP(movie)RV(saw+RP))
61	RV' => PP NP NP *RV	RV'(PP(வீட்டில்)NP(அவனுக்கு)NP(புத்தகத்தைக்)RV(கொடுத்த)) RV'(PP(at home)NP(him)NP(book)RV(gave+RP))
62	RV' => NC *RV'	RV'(NC(படத்தைப் பார்த்து)RV'(கருத்தைச் சொன்ன)) RV'(NC(movie saw+VBP)RV'(comments said+RP))
63	RV' => ADVP *RV'	RV'(ADVP(கவனமாகப்RV'(பார்த்த)) RV'(ADVP(carefully)RV'(saw+RP))
64	RV' => *RV'	RVP(RV'(பார்த்த)) RVP(RV'(saw+RP))
	Relative Clause (RC) Rules	
65	RC => *NP *VP	RC(NP(<trace>)RVP(படத்தைப் பார்த்த)) RC(NP(<trace>)RVP(movie saw+P))
	Sentence (S) Rules	
66	S => *NP *VP	S(NP(mtd)VP(அவன் படத்தைப் பார்த்தான்)) S(NP(he)VP(movie saw))

ஒரு விதியின் தலைப்பு உறுப்பு 'என்ற குறியீட்டால் முன் தொடரப்படும் வாக்கியம் பகுப்படும்போது தலைமை உறுப்பின் இலக்கண பண்புக்கூறுகள் இடப்பக்கமுள்ள இறுதியுறாத ஒன்றிற்குக் கடத்தப்படுகிறது. M என்ற சொல் வகைப்பாடு கூட்டுப்பெயர்களைக் கையாளவேண்டி பயன்படுத்தப்படுகிற தனித்தன்மையான வகைப்பாடாகும். 'M' என்பது மாறுபடுத்தும் தொடரியல் (Modifier Phrase (MP)) தன்மை உறுப்பாகும். கிழமை வேற்றுமைக் குறியீடு உள்ள ஒரு பெயர் மட்டும்தான் கிழமைத்தொடரின் (Genitive Phrase/GP) தலைமையாக

வரும். பின்னுருபுத் தொடர்கள் பெயரைத் தொடர்ந்து முன்,பின் என்ற பின்னுருபுகளால் மெய்ப்படுத்தம் செய்யப்படும். முற்றுப்பெறாத எச்சத்தொடரில் (Non-finite Clause/NC) தலைமை ஒரு முற்றுப்பெறாத வினையாகும். தமிழில் நான்கு வகையான முற்றுப்பெறா வினைகளிலிருந்து நான்கு வகையான முற்றுப்பெறா எச்சத்தொடர்கள் கிடைக்கின்றன. எச்சத்தொடர் (Relative clause (RC)) பெயர்களுக்கு முன்னர் வந்து அவற்றை அடை செய்யும்.

5.6 இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர் பகுத்தல்

இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர் பகுத்தல் (chunking) ஒருவாக்கியத்தின் தனிநிலை பெயர்தொடர்கள், வினைக் குழுமங்கள், பெயரடைத் தொடர்கள், வினையடைத் தொடர்கள் இவற்றை அடையாளங்காணுதலைக் குறிப்பிடும். இது தொடர்களின் எல்லையையும் புலக்குறிப்பையும் கண்டுபிடிப்பதை உட்படுத்தும்.

இடுவரல்-விடுவரல் தனிக்குறிப்பீடுகள்

Input-Output Specification

Input: TKN_, CAT_

Output: ADDR_ (ADDR என்பதன் மதிப்பு மாறுகையில் மீளமைப்புச் செயல்பாடு நடக்கிறது என்பது இதன் அர்த்தமாகும்.)

மேலே தரப்பட்டுள்ள இடுவரல் தனிக்குறிப்பீடுகள் SSF-இல் வருணனை செய்வதற்கு TKN_ மற்றும் CAT_ என்ற தனிப்பண்புகள் அதற்குத் தேவை அல்லது அது பன்படுத்துகின்றது என்று அர்த்தம்தரும். வேறுவிதமாகக் கூறினால் இடுவரல் வாக்கியத்தில் சொற்கள் அல்லது டொக்கன்கள் சொல்வகைப்பாட்டு வகையுடன் தரப்பட்டுள்ளன.

மேலே தரப்பட்டுள்ள விடுவரல் தனிக்குறிப்பீடுகள் 'ADDR_' இந்தத் தொகுதியால் வரையறவிளக்கம் பெறும் என்று அர்த்தம் பெறும்.

எடுத்துக்காட்டு:

வருணனை: தனிநிலைப் பெயர்த்தொடர்கள் தொடர் உறுப்புகளாகக் (NPs) குழுமப்படும். வினைத் தொடர்ச்சிகள் வினைத் தொடர் உறுப்புகளாகக் (VG) குழுமப்படும்.

5.6.1 இ-இ இயந்திர மொழிபெயர்ப்புத் திட்டத்திற்குத் தேர்ந்தெடுக்கப்பட்டுள்ள தொடர்கள்

இதில் தொடர்களின் (chunks) வரையறை விளக்கமும் தொடர்களின் வகைகளும் சில தனித்துவம் வாய்ந்த நேர்வுகளும் பேசப்படும். இந்திய மொழிகளுக்கு இடையிலான மொழிபெயர்ப்புத் திட்டத்தில் 11 தொடர்கள் தெரிந்தெடுக்கப்படு அவை தொடர் உறுப்பு அடையாளப் படுத்தலுக்குப் பயன் படுத்தபடுகின்றன.

முக்கியத் தொடர்களின் குழுமம் அவற்றின் உருப்படுத்தமும்

	முக்கிய தொடர் குழுமம்	**உருப்படுத்தம்**
1	பெயர்த் தொடர்	NP
2	வினைமுற்றுத் தொடர்	VGF
3	வினைமுற்றாத் தொடர்	VGNF
4	வினைப்பெயர்த் தொடர்	VGNN
5	வினை எச்சத்தொடர்	VGINF
6	பெயரைடைத் தொடர்	JJP
7	வினையடைத் தொடர்	RBP

8	இணைப்புத் தொடர்	CCP
9	வியப்புத் தொடர்	INTJ
10	எதிர்மறைத் தொடர்	NEGP
11	வெற்றிடத் தொடர்	BLK

5.6.1.1 பெயர்த் தொடர் (NP chunk)

பெயர்த்தொடர்கள் (NP chunks) NP என அடையாளப் படுத்தப்படும்; இது மறுதரவாக வராத பெயர்த்தொடர்களையும் பின்னுருபுத் தொடர்களையும் உட்படுத்தும். பெயர்த்தொடரின் தலைமை ஒரு பெயராகும். சிறப்பு செய்யும் சொற்கள் பெயர்த்தொடரின் இடது பக்க எல்லையை நிறுவும்; வேற்றுமையோ தலைமைப் பெயரோ வலது பக்க எல்லையை நிறுவும். பெயரை அடைசெய்யும் வருணனைப் பெயரடை(கள்) பெயர்த்தொடரின் பகுதியாக அமையும். தலைமைப் பெயருடன் நங்கூரமிட்டுள்ள குறைச் சொற்களும் பெயர்த்தொடருடன் குழுமப்படும். அது பெயருக்குப் பின்னரோ வேற்றுமைக்குப் பின்னரோ வந்தால், அது தொடரின் வலது பக்க எல்லையை நிர்ணயிக்கும். பெயர்த்தொடர்களின் சில எடுத்துக்காட்டுகள் கீழே தரப்பட்டுள்ளன.

((குழந்தைகள்_NN))_NP, ((சில_QF குழந்தைகள்))_NP, ((சில_QF நல்ல_JJ குழந்தைகள்_NN))_NP, ((மேசை மேல்_PSP))_NP, ((ஒரு_QC கறுத்த_JJ குதிரை_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகம்_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகத்தின்_NN மேல்_PSP))_NP

5.6.1.2 வினைத் தொடர் (VG chunk)

வினைத் தொடர் உறுப்புகள் பலவகைப்படும். ஒரு வினைக் குழுமம் முக்கிய வினையையும் இருப்பின் துணை வினைகளையும் உள்ளடக்கும். பின் வருவது வினைத் தொடர் உறுப்புகளின் எடுத்துக்காட்டுகளாகும்.

((போனான்)), ((போய்விட்டான்)), (போய்க்கொண்டிருக்கிறான்))

5.6.1.3 பெயரைடைத் தொடர் (JJP chunk)

பெயரடைத் தொடர் உறுப்பு JJP என அடையாளப்படுத்தப்படும். இத்தொடர் பயனைலைப் பெயர்தொடர்களையும் உள்ளடக்கி எல்லாப் பெயர்த்தொடர் உறுப்புகளையும் கொண்டிருக்கும். இருப்பினும் பெயருக்கு முன்னர் வரும் பெயரடைகள் பெயர்த் தொடர் உறுப்புடன் சேர்த்துக் குழுமப்படும்.

5.6.1..4 வினையடைத் தொடர் (RBP chunk)

இத்தொடர் உறுப்பின் பெயரும் சொல்வகை அடையாளப் படுத்தலின் போது பயன்படுத்தப்படுகிற சொல்வகை அடையாளத்தின் அடிப்படையில் அமைகின்றது. எல்லா வினையடைத்தொடர் உறுப்பும் RBP என அடையாளப்படுத்தப்படும். இந்தத் தொடர் உறுப்பு எல்லாச் சுத்தமான வினைத்தொடர்களையும் உட்படுத்தும்.

5.6.1.5 இணைப்புத் தொடர் (CCP chunk)

இணைப்புத் தொடர் உறுப்புகள் யாவும் CCP என அடையாளப்படுத்தப்படும். இத்தகவல் பெரிய வாக்கிய அமைபுகளை உருவாக்கத் தேவையாகும்.

5.6.1.6 வியப்புத் தொடர் (INTJ chunk)

எல்லா வியப்புத் தொடர் உறுப்புகளும் INTJ என அடையாளப்படுத்தப்படும்.

5.6.1.7 எதிர்மறைத் தொடர் (NEGP chunk)

வினையைச் சுற்றி இருக்கும் எதிர்மறைக் குறைச்சொல் வினைக் குழுவுடன் குழுமப்படவேண்டும். அது வினையிலிருந்தோ பெயரிலிருந்தோ விலகி இருந்தால் அது தனியான தொடர் உறுப்பாக குழுமப்படவேண்டும்.

5.6.2. தொடர் பகுத்தலின் எடுத்துக்காட்டுகள்

AU-KBC-நிறுவனதில் நடைபெற்றுவரும் இந்திய மொழியிலிருந்து இந்திய மொழிபெயப்பு (தமிழ்-இந்தி) ஒழுங்குமுறையின் இரு கூறுகள் இங்கு எடுத்துக்கொள்ளப்பட்டுளன என்பதன் அடிப்படையில் இவ்விரு செயல்பாடுகளையும் வெளிப்படுத்தும் பொருட்டுத் தமிழ் தரவுத் தொகுதியில் எவ்வாறு சொல்வகை அடையாளப்படுத்தலும் தொடர் பகுத்தலும் நடைபெறுகின்றன என்பதைக் கூறும் முகமாகக் கீழே எடுத்துக்காட்டுகள் தரப்பட்டுள்ளன.

<NP> தலம்/NN :/SYM </NP> <NP> வைணவ/JJ தலம்/NN </NP>
<NP> அமைவிடம்/NN :/SYM </NP> <NP> திருச்சியிலிருந்து/NN </NP> <NP>
3/QC மைல்கள்/NN தூரம்/NN </NP>

<NP> திருக்கோயிலின்/NN பெயர்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC அரங்கநாதசுவாமி/NNP திருக்கோயில்/NN </NP>
<NP> இறைவன்/NN /SYMஇறைவி/NN திருநாமங்கள்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC அரங்கநாத/NNPC சுவாமி/NNP </NP> <NP> ஸ்ரீ/NNC ரங்கநாச்சியார்/NNP </NP>
<NP> தீர்த்தங்கள்/NN :/SYM </NP> <NP> சந்திரபுஷ்கரணி/NN </NP> <CCP> மற்றும்/CC </CCP> <NP> சூரிய/JJ புஷ்கரணி/NN </NP>
<NP> முக்கிய/JJ விழாக்கள்/NN :/SYM </NP> <NP> வைகுண்ட/NNC ஏகாதசி/NN பிரம்மோற்சவம்/NN </NP>
<NP> சிறப்பு/NN அம்சம்/NN :/SYM </NP> <NP> 108/QC வைணவத்/JJ திருத்தலங்களில்/NN </NP> <JJP> முதன்மையான/JJ </JJP>
<VGF>திருக்கோயிலாகும்/VM ./SYM </VGF>
<NP> 12/QC ஆழ்வார்களில்/NN </NP> <NP> 11/QC ஆழ்வார்கள்/NN </NP>
<NP> இதன்/PRPபுகழை/NN </NP> <NP> பெருமையை/NN </NP> <NP> கீர்த்தியைப்/NN </NP> <VGF>பாடியுள்ளார்கள்/VM ./SYM </VGF>
<NP> இந்த/DEMதலத்தில்தான்/NN </NP> <NP> ஆண்டாள்/NNP </NP> <NP> திருப்பாணாழ்வார்/NNP </NP> <NP> பீபிநாச்சியார்/NNP </NP> <CCP> மற்றும்/CC </CCP> <NP> பலர்/NN </NP> <NP> அரங்கனின்/NNP திருவடியை/NN </NP> <VGF>அடைந்தனர்/VM ./SYM </VGF>
<RBP> சுயம்புவாக/RB </RBP> <NP> உருவான/JJ ஒரே/QC திருக்கோயில்/NN ./SYM </NP>
<NP> போக்குவரத்து/NN மையம்/NN :/SYM </NP> <NP> திருச்சி/NNP ./SYM </NP>
<NP> அரங்கன்/NNP </NP> <VGNF> குடிகொண்டுள்ள/VM </VGNF> <NP> திருவரங்கம்/NNP </NP<NP> திருச்சி/NNP /SYM</NP> <NP> சென்னை/NNP கார்டு/NN ரெயில்/NN பாதையில்/NN </NP> <BLK> உள்ள/PSP </BLK> <NP> ரேVயில்/NN </NP> <VGF>நிலையமாகும்/VM ./SYM </VGF>

<NP> சென்னையிலிருந்து/NNP </NP> <NP> 207/QC மைல்கள்/NN தூரத்திலும்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP சந்திப்பிலிருந்து/NN </NP> <NP> 3/QC மைல்கள்/NN தொலைவிலும்/NN </NP> <VGF>உள்ளது/VM ./SYM </VGF>

<NP> திருவரங்கம்/NNP நகரம்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP மாநகரின்/NN புறநகரில்/NN </NP> <BLK> உள்ள/PSP </BLK> <JJP> சிறிய/JJ </JJP> <VGF>நகரமாகும்/VM ./SYM </VGF>

<NP> ஒருபுறம்/NN காவிரியும்/NNP </NP> <NP> மற்றொருபுறம்/NN கொள்ளிடமும்/NNP </NP> <VGNF>பாய்ந்தோடும்/VM </VGNF><NP> இந்த/DEMதிருத்தலத்துக்கு/NN </NP> <RBP> ஆண்டுமுழுவதும்/RB </RBP> <NP> ஆயிரக்கணக்கான/JJ இந்துக்கள்/NN </NP> <VGNF>வந்து/VM </VGNF><NP> அரங்கனை/NNP </NP> <VGF>வழிபட்டு/VM செல்கின்றனர்/VAUX ./SYM </VGF

<RBP> குறிப்பாக/RB </RBP> <NP> வைஷ்ணவர்களுக்கு/NN </NP> <NP> இதுPSP</NP> <JJP> முக்கியமான/JJ </JJP> <VGF>இடமாகும்/VM ./SYM </VGF>

இயல் 6

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை (ILTLMT system)

6.1. அறிமுகம்

6.1.1 நோக்கம்

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்புத் திட்டம் இணையும் அடிப்படையிலான மொழிபெயர்ப்பு திட்டம் 2006 அக்டோபரில் மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில்

நுட்ப அமைச்சின் கீழ் தகவல் தொழில் நுட்பத் துறையின் நிதி நல்கையால் நடைபெற்று வருகின்றது. இதன் முதல் நிலை 2006 அக்டோபர் 30-ஆம் நாள் தொடங்கப்பட்டு 2010 ஏப்பிரலில் முடிவுறது. இத்திட்டம் இந்திய மொழியிலிருந்து மற்றொரு இந்திய மொழிக்கு இயந்திர மொழிபெயர்ப்பை உருவாக்கம் இணையம் அடிப்படையிலான மொழிபெயர்ப்புத் திட்டமாகும். இந்த ஒழுங்குமுறை இரு திசையில் செயல்படும் மற்றும் பொதுவான களத்தில் பின்வரும் மொழிகளுக்கிடையில் செயலாற்றும்.

தமிழ் - இந்தி
தெலுங்கு - இந்தி
மராத்தி - இந்தி
பெங்காளி - இந்தி
தமிழ் - தெலுங்கு
உருது - இந்தி
கன்னடா - இந்தி
பஞ்சாபி - இந்தி
மலையாளம் - தமிழ்

இந்திய மொழி பேசுபவர்கள் இந்த ஒழுங்குமுறையைப் பயன்படுத்துவார்கள். இந்த ஒழுங்குமுறையைக் கற்றல் நுட்பங்களைப் (Machine learning teachiques) பயன்படுத்தி மேலும் இயலும். இந்த ஒழுங்குமுறை இரண்டு குறிப்பிட்ட களங்களுக்கு இயந்திர மொழிபெயர்ப்பை செய்யும்.

பின்வரும் அடிப்படைக் கருவிகள் இந்திய மொழி மொழிப்பெயர்ப்பு ஒழுங்குமுறையின் பகுதியாக உருவாக்கப்பட்டுள்ளது.

1. சொல்வகைப்பாட்டு அடையாளப்படுத்தி (POS Tagger)
2. தொடர் பகுப்பான் (Chunker)

3. உருபனியல் பகுப்பாய்வி (Morph Analyser)

4. வட்டாரச் சொல் குழும (Local word Grouper)

5. மாற்றமைவுப் பகுதி (Transfer Module)

6. இருமொழியப் பொருத்தம் (Billingual Mapping)

7. வாக்கிய நிலை உருவாக்கம் (sentence level generation)

8. வட்டாரச் சொல் குழுமப் பிரிப்பான் (local word group splitter)

9. உருபனியல் உருவாக்கம் (morphological generation)

6.1.2 வாய்ப்பு (scope)

இந்த ஒழுங்குமுறை நிலைப்பேறு பெற்ற மொழியில் எழுதப்பட்ட இணைய பக்கங்கள் அல்லது கால இதழ்கள், நாளிதழ்கள் மீது வேலை செய்யும். சரியான தன்மை பயன்படுத்துவோர் திருப்தி (accuracy / user satisfaction) 85 % - 90 % விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிப்பெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்பட்டுள்ளது மற்றும் 80– 85 % திறமையுடன் பொதுப் பயன்பாடு இயந்திர மொழிபெயர்ப்பு உருவாக்கப்படுள்ளது.

6.1.3. சரியான தன்மை/ பயன்படுதுவோர் திருப்தி (Accuracy/user satisfaction)

85-90 விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிபெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்படும். மற்றும் 80-85 விழுக்காடு திறமையுடன் பொதுப்பயன்பாடு இயந்திர மொழிபெயப்பு உருவாக்கப்படும்.

6.1.4 ஒழுங்குமுறைச் சிறப்பு (system specification)

இந்திய மொழிபெயர்ப்புத் திட்டம் இணையத்துடன் தொடர்பு கொண்டு பயன்படுத்துவோருக்குத் தானியக்க மொழிபெயர்ப்பு சேவைஇயைத் தரவேண்டி ஒரு பரிமாறி (Server) இயங்கும். பின்வருவன மென்பொருள் சிறப்பீடுகளாகும்.

Plat form	:	Fedora Core 4 (Kernel – 2.6.11)
Web server	:	Apache – 2.0
Data base	:	GDBM – 1.8
Compiler	:	gcc 4.0

6.1.5 நெறிமுறை (methodology)

இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறை பல இயற்கைமொழி ஆய்வு ஆய்வுக் குழுவினர் பங்களிப்பால் உருவாக்கப்பட்டுளது. இந்திய மொழி பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கம் பெரிய செயல்பாடு சிறிய செல்பாடுகளாப் பகுக்கப்பட்டுள்ளது. பங்களிக்கும் ஒவ்வொரு குழுவினரும் ஒன்றோ அதற்கு மேலோ வேளைகளை எடுத்துக்கொண்டு இணையாக வேலை செய்துள்ளனர். பெரும்பாலான பகுதிகள் மொழிச் சுதந்திரமான (language independent engine) இயந்திரமாகவும் மொழிச் சிறப்பான தரவாகவும் (language specific data) பகுக்கப்பட்டுள்ளது. இது போல் ஒரு மொழிக் குழுவினர் ஒரு குறிப்பிட்ட மொழிக்குத் தரவைத் தருவர். ஒரே இயந்திரம் பல மொழிகளுக்கு வேலை செய்யும்.

5.1.5.1 பகுதிகளாக அமைத்தல்

இந்திய மொழிபெயர்ப்பு ஒழுங்குமுறை பல பகுதிகளைக் கொண்டது. ஒவ்வொரு பகுதியும் ஒரு தர்க்கம்சார் வேலையைச் செய்யும், பெரும்பாலும் இந்த

வேலை சிறியதாகும். இதன் காரணமாக ஏதாவது மாற்றம் ஏற்பட்டால் சிறிய பகுதியில் அதை எளிதாகச் செய்யவியலும்.

5.1.5.2 சக்தியின் தரமான அமைப்பு (Shakthi Standard Format (SSF)

எல்லாப் பாகங்களும் (modules)வடிவ அமைப்பு நிர்ணயிக்கப்பட்ட தரவின் ஒழுக்கில் செயல்படும். ஒரு பகுதி அதன் வேலையில் வெற்றி பெற்றால் அது வெளிPயீடு ஒழுங்குப் புதிய சின்னத்தையோ ஆய்வையோ சேர்க்கும். பெரும்பாலான பகுதிகளுக்குத் தனிப்பகுதிகளின் வட்டாரச் சிக்கல்களை உறுதியான கட்டுப்பாட்டில் வைக்க இந்த அணுகுமுறை உதவும்.

5.1.5.3 தோல்விகளை நேர்செய்தல்

(SSF) வடிவமைப்பு பகுதியின் ஒவ்வொரு நிலையில் ஏற்படும் தோல்விகளை நேர்செய்ய வசதி செய்யும். எடுத்துக்காட்டாக அதே SSF -இன் மீது அதன் சின்னத்தின் மதிப்பை விளைவிக்கத் தவறினால் கீழ் ஒழுங்குப் பகுதிகள் (down streom modules) தரவின் ஒழுங்கின் மீது தொடர்ந்து செயல்படும்.

5.1.5.4 தெளிவு

இந்தியமொழி இயந்திர மொழிப்பெயர்ப்பில் SSF வடிவமைப்பின் பயன்பாடு ஒவ்வொரு பகுதியின் உள்ளீட்டிற்கும் வெளிப்யீட்டிற்கும் முன் நிகழ்ந்திராதத் தெளப்வைப் பெற உதவுகிறது. ஒரு பகுதியின் உரை சார் வெளிப்யீடு மனிதப் பயன்பாட்டிற்கு மட்டுமல்லாமல் அதன் உள்ளீடாக தரவு ஒழுக்கில் அடுத்து வருகிற பகுதியிலும் பயன்படுத்தப்படும்.

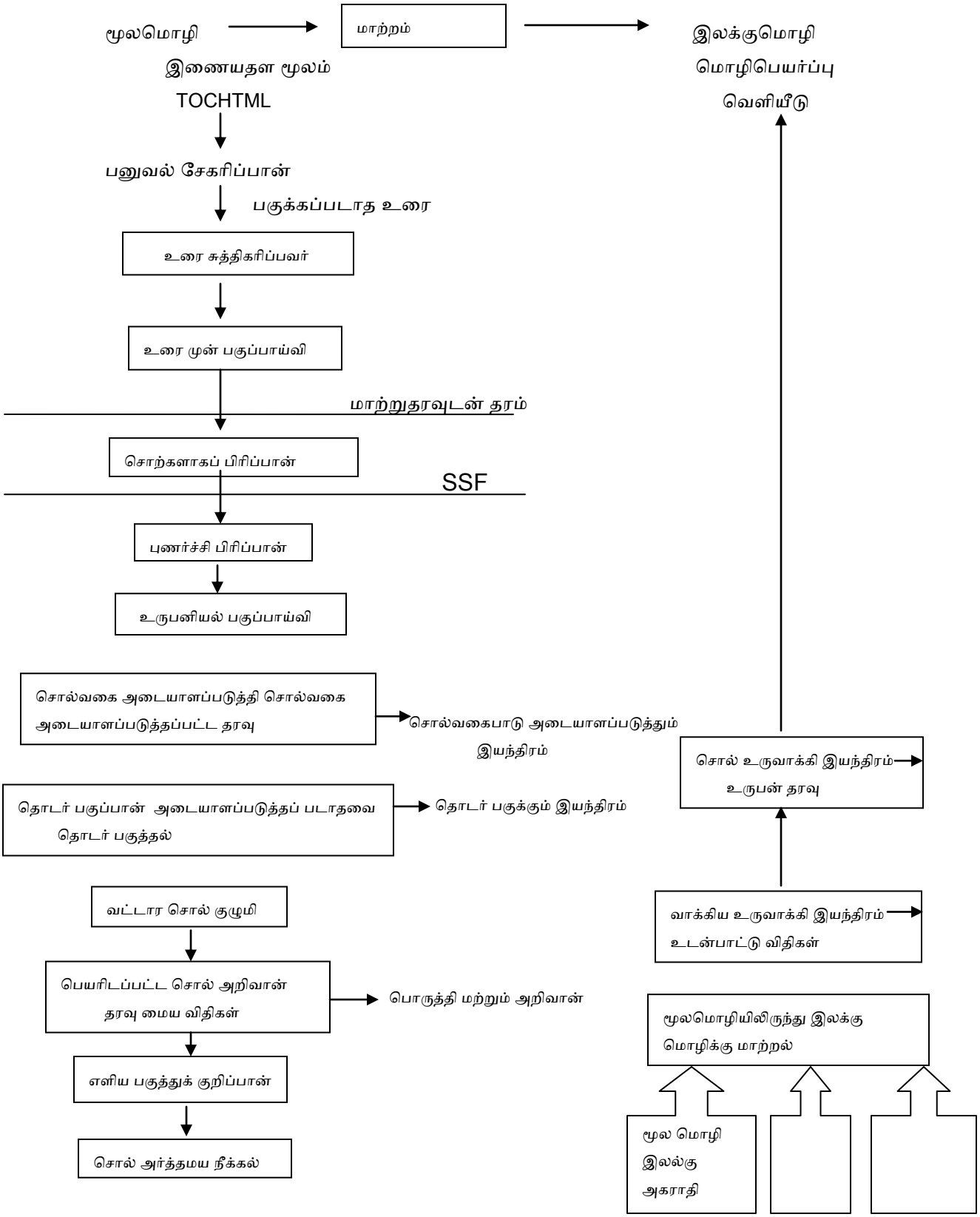
5.1.5.5 டாஷ்போர்டு

இந்திய மொழி மொழிபெயர்ப்பின் எல்லாப் பகுதிகளும் டாஷ்போர்டின் (dash board) உதவியால் இயங்குகிறது. இது குறிப்பிட்டபடி தரவு ஒழுக்கின் பைப்லைன்களை உருவமைப்பு செய்கிறது. இருப்பினும் பைப் லைன்களுக்குப் பதிலாக எல்லாப் பகுதிகளும் செயல்படும் பங்கிடப்பட்ட நினைவகத்தையோ பிளாக் போர்டையோ (Black Board) நிறுவுவதால் ஒங்குமுறையை விரைவுப்படுத்த இயலும். இருப்பினும் இது பங்கிடப்பட்ட நினைவாகத்தில் (அல்லது நினைவக உருப்படுத்தத்தில் SSF) இயங்குவதற்கு வேண்டி பகுதிகள் எழுதப்பட்டிருக்கவேண்டும் என்று அவசியப்படும்.

5.2 முழுவிளக்கம்

இந்திய மொழியிலிருந்து இந்திய மொழிக்கான மொழிபெயர்ப்பு ஒழுங்குமுறை ஆய்தல் - மாற்றியமைத்தல் - உருவாக்குதல் என்ற மேற்கோள் வாய்ப்பாடு அடிப்படையில் செயல்படும். முதலாவது மூலமொழியின் ஆய்வு செய்யப்படும். பின்னர் இலக்குமொழிக்குச் சொற்றொகையும் ஆயப்பட்ட அமைப்பும் மாற்றம் செய்யப்படும். இறுதியாக இலக்கு மொழி உருவாக்கப்படும். இந்திய மொழிகள் ஒற்றுமையுள்ளதாக இருப்பதாலும் இலக்கண அமைப்புகளைப் பங்கீட்டுக் கொள்வதாலும் ஆழமற்ற பகுத்துக் குறித்தல் செய்யப்படும். மாற்றமைவு இலக்கணப் பகுதி எளிமையாக வைக்கப்படும். பொருண்மை களச் சிறப்புப் பெயரிடப்பட்ட சொற்களின் புரிந்து கொள்வான்களாலும் பொருத்தமான அகராதி போன்றவற்றாலும் கையாளப்படும். ஆழமற்ற பகுத்துக் குறித்தலானது, உருபனியல் ஆய்வு, சொல் வகைப்பாடு அடையாளப்படுத்தல் மற்றும் தொடர்பகுத்தல் என்பனவற்றை உள்ளடக்கும். முதலாவது (உருபனியல் பகுத்தாய்வு) விதி அடிப்படையிலானது, இரண்டாவது (சொல் வகைப்பாடு அடையாளப்படுத்தல்) புள்ளியியல் அடிப்படையிலானது, மூன்றாவது (தொடர்ப்பு)

விதி அடிப்படை மற்றும் புள்ளியியல் அடிப்படை இரண்டும் இணைந்தது. முழு ஆய்வு வேலையும் பல பாகங்களைப் பிரிக்கப்பட்டுள்ளது. ஒவ்வொரு பாகமும் சிறிய தர்க்க வேலையைச் செய்யும். இந்த ஒழுங்குமுறையின் முழு அமைப்பும் கீழே தரப்பட்டிருக்கும் படத்தில் காட்டப்பட்டுள்ளது.


மூலமொழி
மாற்றம்
இலக்குமொழி
இணையதள மூலம்
மொழிபெயர்ப்பு
TOCHTML
வெளியீடு
பனுவல் சேகரிப்பான்
பகுக்கப்படாத உரை
உரை சுத்திகரிப்பவர்
உரை முன் பகுப்பாய்வி
மாற்றுதரவுடன் தரம்
சொற்களாகப் பிரிப்பான்
SSF
புணர்ச்சி பிரிப்பான்
உருபனியல் பகுப்பாய்வி
சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தப்பட்ட தரவு
சொல்வகைபாடு அடையாளப்படுத்தும் இயந்திரம்
சொல் உருவாக்கி இயந்திரம்
உருபன் தரவு
தொடர் பகுப்பான் அடையாளப்படுத்தப் படாதவை தொடர் பகுத்தல்
தொடர் பகுக்கும் இயந்திரம்
வட்டார சொல் குழுமி
வாக்கிய உருவாக்கி இயந்திரம்
உடன்பாட்டு விதிகள்
பெயரிடப்பட்ட சொல் அறிவான் தரவு மைய விதிகள்
பொருத்தி மற்றும் அறிவான்
மூலமொழியிலிருந்து இலக்கு மொழிக்கு மாற்றல்
எளிய பகுத்துக் குறிப்பான்
மூல மொழி இலல்கு அகராதி
சொல் அர்த்தமய நீக்கல்

ஒவ்வொரு முக்கியமான பகுதியும் கீழே விளக்கப்பட்டுள்ளன. ஒவ்வொரு பகுதியும் அதற்கு நேரே தரப்பட்டுள்ள முக்கிய வேலையைச் செய்கின்றன. இருப்பினும் முக்கிய வேலை தவிர அவை சில முன்னாய்வையும் (Preprocetsins) பின்னாய்வையும் (Post processing) செய்கின்றன. எடுத்துக்காட்டாகத் தொடர்பகுப்பான் பகுது தொடர் பகுதிகளைக் கண்டுபிடிக்கின்றன. இவ்வாறு தொடர்பகுப்பான் பகுது முக்கிய வேலையான தொடர் பகுத்தலைச் செய்கின்றது. இருப்பினும் தொடர் பகுதியின் தலைமையைக் கண்டுபிடிக்கும் பின்னாய்வைச் செய்கின்றது. மேலும் தலைமைப் பண்புக் கூறுகளைத் தொடர்பகுதியின் பண்புக்கூறுகளாக நகல் செய்கின்றது.

5.2.1 உருபனியல் பகுப்பாய்வு

ஒரு சொல் தரப்படும் போது உருபனியல் ஆய்வி அதன் வேரையும் சொல்லின் இலக்கணப் பண்புக்கூறுகளையும் அடையாளங்கண்டு கொள்கிறது. திரிபு வள்ம் இல்லாத மொழிகளுக்கு எல்லாச் சொல் வடிவுகளையும் கொண்ட எளிய அகராதி நோக்கீடு போதுமானது. தமிழ், தெலுங்கு போன்ற திரிபு வளமுள்ள மொழிகளுக்கு அகராதியை உருவாக்குவது எளிதல்ல. இது பெரிய சேகரிப்பையும் உயர்ந்த செயல்பாட்டுக் கணிப்பையும் வேண்டும். இதற்கு நல்ல மாற்று வழி வேர்ச்சொற்களின் அகராதியை வைத்துக் கொண்டு உருபனியல் ஆய்வியைப் பயன்படுத்தி சொ வடிவை வேராகவும் இலக்கணப் பின்னுருபுகளாகவும் பிரிப்பதாகும்.

மேற்கோள் வாய்ப்பாடு அடிப்படையிலான அணுகுமுறைகள் (Paradigm based approaches) இந்திய மொழிகளுக்கு நல்ல வெற்றி மதிப்பீட்டைத் தருகிறது.

5.2.2 சொல்வகைப்பாடு அடையாளப்படுத்தி

சொல்வகை அடையாளப்படுத்தல் வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டைத் தருவதாகும். வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள் என்று அடையாளம் தருவது ஒரு வாக்கியத்தில் ஒவ்வொரு உறுப்பு பின் ஒரு பங்களிப்பை ஆய்வதற்கு உதவும். சொல்வகைப்பாட்டு அடையாளப் படுத்துவதற்கு விதி அடிப்படை, புள்ளியியல் அடிப்படை போன்ற பல அணுகுமுறைகள் இருக்கின்றது. இங்கு புள்ளியல் அணுகுமுறை பின்பற்றப்படுகின்றது.

5.2.3 தொடர் பகுப்பான்

தொடர்பகுத்தல் (Chunking) ஒரு வாக்கியத்தில் உள்ள எளிய பெயர்த்தொடர்கள், வினைக்குழுமங்கள், பெயரடை தொடர் மற்றும் வினையடை தொடர்களை அடையாளம் காண்பது ஆகும். இது தொடர் பகுது மற்றும் புலக்குறிப்பு இவற்றின் எல்லையை அடையாளங்காண்பதை உட்படுத்தும்.

5.2.4 வட்டாரச் சொல் குழுமி (Local Word Grouper)

வட்டாரச் சொல் குழுமி விபக்தி கணினியக்கத்தின் (Vibhakti Computation) தொழில்நுட்ப வேலையைச் செய்யும் தனிச்சொற்களாக வரும் துணை வினைகள் தனி பொருளடக்கச் சொற்களாகக் கருதப்படுவதில்லை. ஏனென்றால் அவை முதன்மை வினையின் தலைமையாக இருக்கும் தொடர்பகுதியின் இலக்கணப் பண்புக் கூறுகளைச் சிறப்பீடு செய்யும். இது இந்தியில் தனிச் சொற்களாக வருகின்ற பெயர்விபக்தி என்று அழைக்கப்பெறிகின்ற பின்னுருபுகளுக்கும் உண்மையாகும். வட்டாரச் சொல் குழுமி பொருத்தமான கலவை நிலை மாறும்

கூட்டுநிலை சொல் குழும ஆக்க விதிகளைத் திறமையாகக் கையாளும். முக்கியமான வேலை வட்டாரத் தகவல் அடிப்படையில் செயல்பாட்டுச் சொற்களைப் (Functional Words) பொருளடக்கச் சொற்களுடன் (Content Words) குழுமுவதாகும். இந்த வேலையை நிறைவேற்ற பயன்படுத்தப்படும் ஒரு செயல்முறை கிரியரூப்பா அட்டவணைகளைப் (Kiya rupa Charts) பயன்படுத்துவதாகும். இந்த அட்டவணைகள் ஒரு தனி செயல்பாட்டைக் குறிக்கும் தொடர்ச்சியான வினைகளிலிருந்து உருவாக்க வேண்டிய குழுமங்களைச் சிறப்பீடு செய்யும்.

5.2.5 மாற்றல் பகுதி (Transfer Module)

மொழியாய்வில் வாக்கியங்கள் அவற்றின் தொடரமைப்பை அடையாளம் காண்பதற்காகப் பகுத்துக் குறிக்கப்பெறும். இந்திய மொழிகளுக்குள் வேறுபாடுகளை விட ஒற்றுமைகள்தான் அதிகம். எடுத்துக்காட்டாகத் தமிழ், இந்தி மொழியினை முழு பகுத்துக் குறித்தலை வேண்டாது. இத்திட்டத்தில் இயந்திர மொழிபெயர்ப்பு முழு வாக்கியப் பகுத்துக் குறிப்பான் இல்லாது நிறைவேற்றப்படும். மூலமொழியின் அமைப்பிற்கு இலக்குமொழி அமைப்பில் நிகரான அமைப்பு இல்லாவிட்டால் அமைப்பு சார் மாற்றம் தேவைப்படும். ஒரு பகுதி பகுத்துக் குறித்தல் அல்லது ஆழமற்ற பகுத்துக் குறித்தல் மாற்றாத்திற்கு உட்படுத்தப்படும் வாக்கியங்களின் குறிப்பிட்ட உறுப்புகளை அடையாளம் கண்டால் போதுமானது.

மொழிகளுக்கிடையில் உள்ள தொடரியல் வேறுபாடு கலவை வாக்கிய கட்டுமானங்களில், குறிப்பாக எச்சத்தொடர்களில் காணப்படும். இது மாற்றமைவு இலக்கணத்தால் இணைக்கப்படும். இந்தப் பகுது எழுத்துருவாக்கத்தின் வேலையையும் உட்படுத்தும். இந்திய மொழிகளுக்குள் (உருது மொழியை உள்ளடக்கி) எழுத்துருவாக்கத்தை நிறைவேற்றும் ஒரு பகுதி உருவாக்கப்பட

வேண்டும். எழுத்துருவாக்கம் சொல்லையோ சொற்களையோ படிப்பவரின் எழுத்துக்கு மாற்றுகிறது. எடுத்துக்காட்டாக ஒருவருக்கு இந்து தெரிந்திருந்தால் தேவநகரியில் வங்காள உரையைப் படிக்க இயலும் என்றால் அதன் பொருளின் சில பாகங்களை அவரால் புரிந்து கொள்ள முடியும். மொழிபெயர்ப்பு ஒரு சொல்லுக்கோ ஒரு தொடர் பகுதிக்கோ தோற்றுப்போனால் எழுத்துருவாக்கம் படிப்பவரை படிக்கவும் புரிந்து கொள்ளவும் அனுமதிக்கும். இந்திய மொழிகள் பல எண்ணிக்கையிலான சொற்களைப் பங்கிட்டுக் கொண்டுள்ளன. எனவே படிப்பவருக்கு ஏற்றவாறு எழுத்துருமாற்றம் செய்தால் அவரால் மூலமொழியில் உள்ள உரையிலிருந்து சிலவற்றைப் புரிந்து கொள்ள இயலும்.

5.2.6 இருமொழிய அகராதி பொருத்தம் (Bilingual Mapping)

உருபனியல் பகுப்பாய்வால் அடையாளம் காணப்பட்ட வேர்ச்சொல் இலக்குமொழி நிகரன்களுக்கு வேண்டி இருமொழி அகராதியில் பார்க்கப்படும். இந்த இருமொழி அகராதி இலக்கு மொழியின் வேர்ச்சொல் நிகரன்களையும் அதன் வகைப்பாடு (பெயர், வினை போன்றவை) மேற்கோள் வாய்ப்பாடு மற்றும் பிற தேவையான தகவல்களைக் கொண்டிருக்கும். இந்த நிலை மூலமொழி இலக்கண மின்னுருபுகளுக்கு இலக்கு மொழி நிகரன்களை அடையாளம் காண அகராதியைப் பயன்படுத்துகிறது.

5.2.7 வாக்கியநிலை உருவாக்கம்

இந்தச் செயல்பாடு தேவையானால் தொடர்பான்களை முதன்மையாக நிரல் மாற்றியமைக்கும். சில வாக்கியநிலை நோக்குகளும் இந்தப் பகுதியால் கையாளப்படும். எடுத்துக்காட்டாக பெயர், வினை உடன்பாடு போன்றவை இங்கு கையாளப்படும். இந்த மட்டம் தொடர்பகுதி மட்ட ஆய்வாகும். வட்டாரச்சொல் வகுப்புகளைக் கொண்ட தொடர் பகுதிகளுக்கு பொருத்தமான

சொற்கள் (அப்படியிருந்தால்) உருவாக்கப்படும். முதன்மை வினை, துணை வினை உடன்பாடு போன்ற உள் தொடர் பகுதி உடன்பாடு இங்கு காணப்படும்.

5.2.8 வட்டாரச்சொல் குழு பிரிப்பான்

இந்தப் பகுதி ஒரு சொல்லுக்கு விபக்தி மற்றும் பிற தகவல்களைக் கணிக்கும் மற்றும் அவற்றைத் தனிச்சொற்களாகத் திருப்பி அனுப்பும்.

5.2.9 உருபனியல் உருவாக்கம்

ஒரு குறிப்பிட்ட இலக்கண வகைக்காக ஒரு சொல்லை உருவாக்க வேர்ச்சொல்லுடன் பொருத்தமான பின்னொட்டுகள் இணைக்கப்பட வேண்டும். உருவாக்கி மொழியின் உருபனியலுக்குத் தக்கவாறு வேர்ச்சொல்லை திரிபுற செய்யும் மற்றும் இலக்குமொழி சொல் வடிவை வெளியிடும். இவ்வாறு உருவாக்கப்பட்ட சொற்கள் முழு இலக்குமொழி வாக்கியத்தை உருவாக்க இணைக்கப்படும்.

5.2.10 மதிப்பீடு செய்தல்

மதிப்பீடு செய்யும் இயக்க நுட்பம் தனி வாக்கியங்களுக்கும் உரைகளுக்கும் வேண்டி இறுதி பயன்பாட்டாளர்களுக்காக நிறுவப்படும். இது புரிந்து கொள்ளுதல் மற்றும் மொழிபெயர்ப்பு தரம் இவற்றிற்காக மதிப்பீடு செய்யப்படும்.

இரண்டாவது நிலையில் மதிப்பீடு ஒழுங்குமுறையின் வேகத்திற்காகவும் பயன்பாட்டிற்காகவும் செய்யப்படும்.

5.3 முடிவுரை

இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் முக்கியமான ஒரு நடைமுறை செயல்பாடாக மாறும் நிலையை நோக்கி வளர்ந்து கொண்டிருக்கிறது. இணைய தளத்தில் ஏராளமான தகவல்கள் தமிழிலும் பிறமொழிகளிலும் இருக்கின்றன. குறிப்பாக ஆங்கிலத்தில் பல முக்கியமான தகவல்கள் இணைய் தளத்தில் வலம் வருகின்றன. இத்தகவல்களை தமிழ் மட்டும் அறிந்த நம்மால் புரிந்து கொள்ள இயலாது. இதற்கு வேண்டி உடனடி மொழிபெயர்ப்பு தேவை. இணைய தளத்தில் இருக்கும் ஆங்கிலத் தகவல்களை உடனடியாக இயந்திரத்தினாலேயே தமிழில் மொழிபெயர்ப்பு செய்து கிடைக்கப் பெற்மானால் அது மிகுந்த வரபிரசாதமாகும். அத்தகைய திசை நோக்கித்தான் தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகிறது. தமிழிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து தமிழுக்கும் தமிழிலிருந்து பிற மொழிகளுக்கும் தமிழ்மொழிக்கும் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகின்றன. தமிழ்ப் பல்கலைக்கழகம், அண்ணா பல்கலைக்கழகம், சென்னைப் பல்கலைக்கழகம், AUKBC நிறுவனம் தமிழ் இணைய பல்கலைக்கழகம் என்பன தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் ஈடுபட்டுள்ளன. இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technological Development Indian Languages) என்ற திட்டத்தின் கீழ் இந்திய மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில்நுட்ப அமைச்சின் (Ministry Communication and Information Technology) கீழ்வரும் தகவல் தொழில்நுட்ப துறையால் பொருளாதார உதவிப்பெற்று பல நிறுவனங்களும் பல்கலைக்கழகங்களும் ஆங்கில மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு இந்திய-இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு என்பனவற்றில் முழுமூச்சுடன் ஈடுபட்டுள்ளன. இம்முயற்சிகள் யாவும் இணையத் தளத்திலிருந்து மொழிபெயர்ப்பை இலக்காகக் கொண்டு அமைகிறது. இந்திய மொழிகளின் மைய நிறுவனம் இந்திய

மொழிகளின் பேச்சு மற்றும் எழுத்து விரிதரவுகளை திரட்டும் திட்டத்தை பெற்று செயல்படுத்தி வருகிறது.

முற்றிலும் தானியக்க மொழியியல் பெயர்ப்பு இன்றைய காலகட்டத்தில் இயலாத வகையால் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பை ஊக்குவிக்கிறது. பொதுவான மொழிக்கு அதாவது பொதுவான மொழி உரைக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவது கடினமாகையால் ஒரு குறிப்பிட்ட பொருண்மைகளை வைத்துக் கொண்டு இயந்திர மொழிபெயர்ப்புச் செய்யும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. ஆங்கில-இந்திய மொழிகள், இந்திய மொழிகள்-இந்திய மொழிகள் மொழிபெயர்ப்பு திட்டத்தில் சுற்றுலா மற்றும் உடல் நலம் (Health) என்பன பொருண்மை களங்களாக ஏற்கப்பட்டு மொழிபெயர்ப்பு செய்யும் முயற்சி மேற்கொள்ளப்பட்டு வருகின்றன.

இயந்திர மொழிபெயர்ப்பு என்பது காலத்தின் கட்டாயமாகும். தமிழுக்கான தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் நாம் ஈடுபட்டே ஆக வேண்டும். இல்லாவிட்டால் இருமொழிய (Bilingual) (ஆங்கிலம், தமிழ்) நிலையிலிருந்து ஒரு மொழிய நிலைக்கு (Mondingual)மாறிவருகின்ற நம் மாணவ சமுதாயத்தைக் காப்பாற்ற இயலாது போய்விடும். இணைய தளத்தில் கிடைக்கும் அறிவு மூலங்களைத் தமிழர்களாய நாம் பெறாது போய் விடுவோம். இதுபோல தமிழில் கிடைக்கும் அரிய தகவல்கள் பிற மொழியினவர்களுக்குப் போய்ச் சேராது. தகவல் பன்மடங்காகப் பெருகிவரும் இக்கால கட்டத்தில் தகவல் பரிமாற்றத்திற்கு இயந்திர மொழிபெயர்ப்பை விடையாகும்.

இயல் 7

முடிவுரை

இயந்திர மொழிபெயர்ப்புக் கோட்பாடுகளையும் கொள்கைகளையும் தமிழில் தருவதும் தமிழுக்கென்று தரவுத்தொகுத் தயாரிப்பது எவ்வாறு அத்தரவுத் தொகுதியை இயந்திரமொழிபெயர்ப்புச் செயல்பாடுகளுக்கு வேண்டி எவ்வாறு அடையாளப்படுத்துவது அதில் வரும் சிக்கல்கள் என்ன அதற்காகன் தீர்வுகள் என்ன என்பது பற்றி விளக்குவதும் தான் இவ்வேட்டின் முக்கிய நோக்கமாகும். தரவுத்தொகுதி தயாரித்தல் மிக முக்கியமான செயல்பாடாக மாறியுள்ளது. இயற்கைமொழி ஆய்விற்கும் இயந்திர மொழிபெயர்ப்பு ஆய்வுக்கும் தரவுத்தொகுதி முக்கியமான பகுதியாக அமைகிறது. ஒரு காலகட்டத்தில் மொழியியலில் கள ஆய்வுகள் மூலம் தரவுகள் திரட்டப்பட்டு, ஆராயப்பட்டு விதிகள் உருவாக்கப்பட்டன. இவ்விதிகள் குறைந்த அளவிலான தரவின் அடிப்படையில் அமைந்து முழுவதுமாக மொழி அமைப்பை விளக்காமல் போய்விட்டதன் காரணமாக இக்கள ஆய்வு குறைகள் நிறைந்ததாய்க் கருதப்படல்லாயிற்று. இதற்கு மாறாகச் சாம்ஸ்கி என்பவர் அறிதிறனை அடிப்படையாகக் கொண்டு மொழியியல் விதிகள் அமைய வேண்டும் என்று வலியுறுத்திக் கூறினார். இந்த இருவித அணுகுமுறையிலும் சிக்கல்கள் இருந்துவந்தன. மொழியை ஒட்டுமொத்தமாக விளக்க கள ஆய்விலிருந்து பெறும் தரவோ அறிதிறன் அடிப்படையில் அமையும் தீர்மானங்களோ போதாது என்பதன் அடிப்படையில் புள்ளியியலைச் சார்ந்து விதிகளை அல்லது தீர்மானங்களை எடுக்கும் நோக்கு முக்கியத்துவம் பெற்றது. இதன் காரணமாக மொழியியல் ஆய்வானது களத் தரவு ஆய்விலிருந்து தரவுத்தொகுதி ஆய்வுக்கு நீட்சியுற்றது. புள்ளியியல் அடிப்படையிலான தீர்மானங்கள் நம்பத்தகுந்ததாக அமைந்தால் தரவுத்தொகுதி தயாரித்தல் மிக முக்கியமான செயல்பாடாக அமைந்தது. இயற்கைமொழி ஆய்வின் எந்தச் செயல்பாட்டிற்கும் தரவுத்தொகுதி அடிப்படையிலான புள்ளியியல் ஆய்வு

முக்கியமாகக் கருதப்பெற்றது. தமிழும் இம்முயற்சிகளிலிருந்து பின்வாங்கவில்லை. தமிழுக்கான தரவுத்தொகுதிகள் பல தயாரிக்கப்பட்டு இன்று பயன்பாட்டில் உள்ளன.

வேறுபட்ட வகையிலான அகராதிகள் தயாரிப்பதற்கும் சொற்களஞ்சியம் தயாரிப்பதற்கும் இயந்திர மொழிபெயர்ப்புக்கும் தரவுத்தொகுதி இன்றியமையாததாகும். இயந்திர மொழிபெயர்ப்புக்கு மிக முக்கியமாகக் கருதப்படும் சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை தரவுத்தொகுதியின் புள்ளியியல் விவரங்களின் அடிப்படையிலானதாகும். பேச்சு உரை மாற்றத்திற்கும் உரை பேச்சு மாற்றத்திற்கும் பேச்சு தரவுத்தொகுதிகள் கைக்கொடுக்கும். இவ்வாறு தரவுத் தொகுதியின் முக்கியத்துவம் குறித்து கூறிக்கொண்டே செல்லலாம்.

அடையாளப்படுத்தப்படாத தரவுத்தொகுதியை வைத்துக்கொண்டு எந்த இயற்கை மொழி ஆய்வையும் செய்ய இயலாது. தரவுத்தொகுதி சொல்வகைப்பாட்டிற்காக அடையாளப்படுத்தப்பட்டால்தான் அத்தரவுத்தொகுதி பயனுள்ளதாக அமையும். தமிழைப் பொறுத்தவரையில் பல சொல் வகைப்பாடுகள் இருப்பதைப் பார்த்தோம். தமிழ்த் தரவுத்தொகுதியானது சொல்வகைப்பாடுகளுக்காக அடையாளப்படுத்தப்பட்டால்தான் அவற்றைப் பயன்படுத்த முடியும். பலவகை சொல்வகைப்பாட்டு குழுமங்கள் வழக்கத்திலிருந்து வருகின்றன. ஆங்கிலத்திற்கென்று உருவாக்கப்பட்ட சொல்வகைப்பாட்டு குழுமங்களைப் பயன்படுத்துவது, இந்திக்கென்று உருவாக்கப்பட்ட சொல்வகைப்பாட்டு குழுமத்தைப் பயன்படுத்துவது, தமிழுக்கென்று தனி சொல்வகைப்பாட்டு குழுமத்தை உருவாக்கிப் பயன்படுத்துவது போன்ற செயல்பாடுகள் இன்று நடைமுறையில் உள்ளன. இந்திய மொழிகளுக்காகச் சொல்வகைப்பாட்டுக் குழுமத்தை IIIT ஹைதராபாத் நிறுவனம் உருவாக்கிப் பயன்படுத்தி வருகின்றது. இந்திய மொழிகளுக்கு இடையிலான

மொழிபெயர்ப்பில் இந்நிறுவனத்தின் சொல்வகைப்பாட்டு அடையாளக் குழுமம் பயன்படுத்தப்பட்டு வருகிறது. இது குறைபாடு உள்ளதாகத் தோன்றுகிறது. இச்சொல்வகைப்பாட்டு அடையாளக் குழுமம் இந்தி மொழியின் அடிப்படையில் உருவாக்கப்பட்டதால் தமிழ் போன்ற திராவிட மொழிகளுக்கு சரியாகப் பொருந்தவில்லை. மைசூரில் உள்ள இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கான குறிப்பாக இந்தி மொழிக்கான சொல்வகை அடையாளக் குழுமத்தை உருவாக்கிப் பயன்படுத்தி வருகிறது. சென்னையிலுள்ள AUKBC நிறுவனம் தமிழுக்கென்று தனியாகச் சொல்வகைப்பாட்டு அடையாளக் குழுமத்தை உருவாக்கிப் பயன்படுத்தி வருகிறது. இந்திய மொழிகளுக்கான சொல்வகை அடையாளக் குழுமத்தை Eagles குறிப்பிடும் பரிந்துரைகளின் அடிப்படையில் தரப்படுத்தும் முயற்சி பெங்களூரில் உள்ள MSRI நிறுவனத்தால் மேற்கொள்ளப்பட்டு வருகிறது. இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்புத் திட்டத்தின் முதல் கட்டம் முடிவுற்ற நிலையில் அத்திட்டத்தில் பயன்படுத்திவந்த சொல்வகைபாட்டுக் குழுமத்தில் மாறுதல் செய்து புதிய சொல்வகைப்பட்டுக் குழுமத்தை உருவாக்கும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன.

இவ்வேட்டில் விளக்கப்படும் மூன்றாவது சிக்கல் தொடர் பகுத்தலாகும். சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி தொடர்களுக்காக அடையாளப்படுத்தப்பட வேண்டும். தொடர்களைக் கண்டறிந்த பின்னர்தான் நாம் வாக்கியதில் இத்தொடர்களின் செயல்பாட்டுகளை உணர்ந்து மூல மொழியை இலக்கு மொழிக்கு மாற்ற இயலும். இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்புத் திட்டந்தில் இதற்கென்று தொடர்கள் நிறுவப்பெற்று புள்ளியியல் அடிப்படையில் தரவுத்தொகுதியை அடையாளப்படுத்தும் செயல்பாடும் நிர்ணயிக்கப்பட்டுள்ளது. ஆழமில்லாப் பகுத்தலின் ஒரு பகுதியாக அமையும் தொடர் பகுத்தல் முழுத் தொடரியல் பகுப்பாய்வைக் காட்டிலும் சிறந்ததாகக் கருதப்படுகின்றது.

இறுதியில் இவ்வேட்டின் நோக்கமாக அமையும் சொல்வகை அடையாளப்படுத்தலையும் தொடர் பகுத்தலையும் உள்ளடக்கிய இந்திய மொழியிலிருந்து இந்திய மொழிக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை பற்றிய விரிவான விளக்கம் தரப்பட்டுள்ளது.

.

துணைநூற்பட்டியல்

Aarts.J. and Meijs, W. (Eds.)1984. Corpus Linguistics: Recent Development in the use of Computer Corpora in English Language Research. Amsterdom-Atlanta. G.A.:Rodopi.

Aarts.J. and Meijs, W. (Eds.)1986. Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora. Amsterdam-Atlanta, G.A.: Rodopi.

Aijmer, K. and Alwood, J. (Eds.) 2004. Dialogic Analysis: New Trends in Dialogue Analysis, Tubingen: Niemeyer.

Aijmer, K. and Altenberg, B. (Eds.) 1991. English Corpus Linguistics: Studies in Honour of Jab Stvartvik. London: Longman.

Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai. 2006. AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages, Technical Report, Language Technologies Research Centre IIIT, Hyderabad.

Arulmozhi. P, Sobha L. 2006. A Hybrid POS Tagger for a Relatively Free Word Order Language. In proceedings of MSPIL-2006, Indian Institute of Technology, Bombay.

Atwell, E. (Eds.) 1993. Corpus Based Computational Linguistics. Amsterdam: Rodopi.

AU-KBC Post Tagset Tamil. AU-KBC Research Centre, MIT Campus of Anna University, Chrompet, Chennai-44.

Bernadini, S. 2000. Competence, Capacity, Corpora, Bologna: CLUEB.

Biber, D. Conrad, S. and Reppen, R. 1998. Corpus Linguistics-Investigating Language Structure and use. Cambridge University Press.

Chomsky, A.N. 1968. Language and Mind. New York: Harcourt Brace.

Chomsky, A.N. 1972. Studies in Generative Grammar. The Hague: Mouton.

Clear, Jereny, 1992. Corpus sampling in New Directions in English Language Corpora (ed.) Gerhars Leitner. Mouton de Gruyter, New york.

CIIL, HINDI, CORPORA Tagset. Finalized on August 29, 2006

Dash, N.S. 2005. Corpus Linguistics and Language Technology (with reference to Indian Languages). New Delhi: Mitial Publications.

Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S, 2008. ”Tamil Part-of-Speech tagger based on SVMTool“, In Proceedings of the COLIPS International Conference on natural language processing (IALP), Chiang Mai, Thailand. 2008. P.No:

Dhanalakshmi V, Anand kumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S. 2008. “Tamil Part-of-Speech tagger based on SVMTool”, Proceedings of International Conference on Asian Language Processing 2008 (IALP 2008), November 2008, Chiang Mai, Thailand.

Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., Rajendran S., 2009 “Chunker For Tamil Using Machine Learning” 7th International Conference on Natural Language Processing 2009 (ICON2009), IIIT Hyderabad, India.December 2009. P.No:

Dhanalakshmi.V, Anand kumar M, Rekha R U, Arun kumar C, Soman K P, Rajendran S. “Morphological Analyzer For Agglutinative Languages Using

Machine Learning Approaches", Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom 2009, Oct 2009 Kottayam, India .

Dhanalakshmi.V , Anand kumar M, Padmavathy P, Soman K P, Rajendran S, 2009 "Chunker For Tamil" Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom 2009, October 2009, Kottayam, India.

Dhanalakshmi. V , Anand kumar M, Soman K P, Rajendran S , 2009 "Postagger And Chunker For Tamil Language" Proceedings of TAMIL INTERNET CONFERENCE 2009 , October 2009, Cologne, Germany.

Anand kumar M, Dhanalakshmi V, Soman K P, Rajendran S. 2009. "A Novel Approach For Tamil Morphological Analyzer", Proceedings of TAMIL INTERNET CONFERENCE 2009, October 2009, Cologne, Germany.

Dhanalakshmi V., Anand Kumar M., Rekha R.U., Arun Kumar C., Soman K.P., Rajendran S., 2009."Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009,IEEE Press, doi: 10.1109/ARTCom.2009. P.No:184

Dhanalakshmi.V , Padmavathy P, Anand kumar M, Soman K P, Rajendran S, 2009 "Chunker For Tamil Using Machine Learning" 7th International Conference on Natural Language Processing 2009(ICON2009), December 2009, IIIT Hyderabad, India.

Dhanalakshmi V, Anand kumar M, Shivapratap G, Soman K.P, Rajendran S. 2009. "Tamil POS Tagging using Linear Programming" International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May-2009. ISSN 1797-9617.

Dhanalakshmi.V , Anand kumar M, Padmavathy P, Soman K P, Rajendran S. 2009. "Chunker For Tamil", Proceedings of International Conference on

Advances in Recent Technologies in Communication and Computing, ARTCom 2009, October 2009, Kottayam, India.

Eagles. Expert Advisory Group on Language Engineering Standards.

Eagles –Recommendations for the Morphosyntatic Annotaton of Corpora. Eagles Document EAG – TCWG-MACIR. Version of Mar, 1996.

Francis, W.N. and Kucera, H. 1964. Manual of Information to accompany A Standard Corpous of present day edited American English. Department of Linguistics, Brown Uniersity, USA.

Garside, R.Leech, G. and Mcenery, A/ (Eds.) 1997. Corpus Annotation Linguistics Information from computer text Corpora.London: Longman.

Gerbig, A.1997. Lexical and Grammatical Variation in a Corpus: A computer Assisted study of Discourse on the Environment. London: Peter Long Publishing.

Ghadassy, M., Henry.A., and Roseberry, R.L. (Eds) 2001. Small Corpus studies and ELT : Theory and practice Amsterdam / Philadelphia. John Benjamins.

Graf,D. 1996. Relative clauses in their Discourse Context : A Corpus – Based study Unpublished M.A., Thesis : Freiburg.

Granger, S. and Tyson, S.P. (Eds) 2003. Extending the scope of corpus-based Research: New Applications. New challenges. Amserdam: Rodopi.

Granger,S. Hung, J.And Tyson, S.P.Eds. 2002. Computer Learner Corpora, Second Language Acquistion and Foreign Language Teaching. Amsterdam: John Benjamins.

Greenbaun, S. (Ed.) 1996. Camparing English Worldwide: the International Corpus of English. Oxford: Clarendon.

Haan, Peter de, 1992. The optimum Corpus sample size? in New directions in English Language Corpora (ed.,) Mouton de Gruyter, New York.

Hunston, S. 2002, Corpora in Applied Linguistics, Cambridge University Press.

ILMT Consortium. 2007, 2008. ILMT System: Software Requirement Specifications (SRS). IIIT Hyderabad.

Jayaram, B.D. 1996. Development of corpora in Indian languages: Problems and suggested solutions. Paper presented at workshop on Indian Language Corpus and its applications at CIIL, Mysore.

Jayaram, B.D Umarani P. 1998. Grammatical Category Disambiguation: A Probabilistic Model in South Asian Language Review, Vol. vII. No.1. Jayaram, B.D and Rajyashree, K.S corpora in Indian Language. Tamil Text

Johansson, S. and Hofland, K. (Eds) 1982. Computer corpora in English Language Research. Bergen: Nowegian computing Center for the Humanities.

Johansson, S. and Stenstrom, A.B. (Eds.) 1991. English computer Corpora: Selected Paper and Research Guide. Berlin: Mouton de Gruyter.

Kennedy, G. 1998. An Introduction to Corpus Linguistics. New York: Addison-Wesley Longman Inc.

Kettemann, C.B. and Marko, G. (Eds.) 2002. Teaching and Learning by Doing Corpus Analysis. Language and Computers: Studies in Practical Linguistics 42. Amsterdam-Atlanta G.A. Rodopi.

Kirk, J.M. (Ed.) 2000. Corpora Galore: Analysis and Techniques in Describing English, Amserdam, Atlanda, G.A. Rodopi.

Kyto, M. Thalainen, O.and Rissanen, M. (Eds.) 1998. Corpus Linguitics, hard and Soft : Proceedings of the 8 th International Conference on English Language Research on Computerised Corpora. Amsterdam: Rodopi.

Kumar, R. and S. Viswanathan, "tamil Morphological Analyser" at http:/www.av. KBC.org research-areas/projects/documentation.html (17.9.2001)

Lancushine, I.E Carol and C.F. Meyer (Eds.) 1997. Synchronic Corpus Linguistics. Bergen, Norway: ICAME.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17 th International Conference on English Language Research Based on Computerized Corpora. Amsterdam:Rodopi.

Lehmann. 1989. the Grammar of Modern Tamil. Pondicherry Institute of Linguistics and Culture. Pondicherry.

Mathhews, P.H. 1972. Inflectional Morphology. Cambridge University Press, London.

Mair, C. and hundt, M. (Eds.) 2000. Corpus Linguistics and Linguistics Theory. Amsterdam-Atlanda, GA: Rodopi.

Mecnery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburg: Edinburgh University Press.

Mcenery, T. Rayson, P.and Wilson, A. (Eds.) 2002. A Rainbow of Corpora: Corpus Linguistics and the Language of the World. Munchen: Licolm Europe.

Meyer, C.F. 2002 English Corpus Linguistics. Cambridge University Press.

Nelson, G., Wallis, S. and Aarts, B. 2002 Exploring Natural Language: Working With the British component of the International Corpus of English. Amsterdam/Philadelphic John Benjamins.

Ookes, M.p. 1998. Statistics for Corpus linguistics. Edinburgh: Edinburg UniversityPress.

Ooi, V.B.V.1997. Computer Corpus Lexicogrphy. Edinburgh: Edinburg UniversityPress.

Oostdijk, N.and Dehann, P. (Eds.) 1994. Corpus Based Research into Language Amsterdam-Atlanda, GA: Rodopi

Partington, A. 1998. Patterns and meanings-Using Corpora for English Language Reseach and Teaching. Amsterdam, Philadelphia: John BEnjamins.

Percy, C., Meyer, C.F., and Lancashire, I. (Eds.) 1996. Synchronic Corpus Linguistics, Amsterdam – Atlanta, G.A: Rodopi.

Peters, B.P., Collins, P. and Smith, A. (Eds.) 2002. New Frontiers of Corpus Research Language and Computers. Amsterdam – Atlanta, G.A. Rodopi.

Rajendran, S. 1995. “A Machine Bridge for Tamil.” Proceedings of Eighth World Tamil Conference. Thanjavur.

Rajendran, S. 1997. “Grammatical Formalism and Computational Analysis of Nominal Compounds in Tamil.” South Asian Language Review, vol 7, no.1., 27-46.

Rajendran.S. 1999. “Spell and Grammer Checker for Tamil.” Paper read in 27th All Indian Conference of Dravidian Lingustics.

Rajendran.S. 2000. “Preliminaries to the Preparation of spell and Grammer checker for Tamil.” Language in Indian WWW.languageindia.com.

Rajendran.S. 2006. “Parsing in Tamil: Present state of Art.” Indian Linguistics: 67:159 – 67.

Rajendran, S., Arulmozi, S.Ramesh Kumar and S.Viswanathan. 2003. “Computational Morphology of Verbal Complex”. In B.Ramakrishna Reddy (ed.) Word Structure in Dravidian. Kuppam: Dravidian University 376 – 398.

Rajendran, S. et al. 2003. “Computational Morphology of Verbal Complex.” In: B. Ramakrishna Reddy (edited) Word Structure in Dravidian, Kuppam: Dravidian University, & Language in India 3:4, www.languageinindia.com, April 2003.

Ramaswamy, V. 2002. Morphological Generator for Tamil. Unpublished M.Phil Dissertation, University of Hyderabad, Hyderabad.

Ranganathan, 1997. “A Lexical phonology Approach Tamil word by Computer”. International Journal of Dravidian Linguistics, 26, 1:57 – 70.

Rekha, R U Anand kumar M, Dhanalakshmi.V , Soman K P, Rajendran S, 2010 “Morphological generator for Tamil a new data driven approach”, Tamil Internet Conference 2010, June 2010, Cemmozhi maanaadu, Coimbatore, India.

Sengupta, G. 1997. “Three models of Morphological processing” South Asian Language Review, 8:1, 1-26.

Shanmugam, C. 2001. “Computer Analysis of Simple sentence in Tamil”. Paper read in UGC SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, AnnamalaiNagar.

Shanmugam, C. 2002. “Grammer and Parser: A program for Syntactic Parsing in Tamil” International Seminar on Tamil Computing 27 -28 February and March 1, 2002. University of Madras: Chennai.

Sivashanmugam, C. 1998. “Language Technology and Morphological Processing”. The Administrator, 45:3.

Sivashanmugam, C. 2000. “A Model for computer Analysis of Verb in Tamil” In working papers in Linguistics, Department of Linguistics, Bharathiyar University, Coimbatore.

Sivashanmugam, C. 2002. “Morphological: processor for Negative Constructions in Tamil” Indian Conference on Natural Language processing, Anna University, Chennai.

Selkairk, E.O. 1983. The Syntax of words. Cambridge, Mass: MIT Press.

Selting, M.and Couper. Kuhlen E. (Eds.) 2001. Studies in International Linguistics. Amsterdam / Philadelphia: John Benjamins.

Sharman, R. 1990. Hidden Markov Model Methods for word Tagging. Report 214. Winchester : IBM UK Scientific centre.

Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford University Press

Souter, C. and Atwell, E. (Eds.) 1993, Corpus based computational Linguistics. Amsterdam: Rodopi.

Stenstorm, A.B. Anderson, G.and Hasund, I.K.2002. Trends n Teenage Talk: Corpus Compilation, Analysis and Findings, Amsterdam: John Benjamins.

Stubbs, M. 1996. Text and corpus Analysis : Computer Assisted Studies of Lanaguage and culture. Oxford : Balckwell Publishers.

Summers, D. 1991. Longman/Lancaster English Language Corpus: criteria and Design Harlow: Longman.

Svartvik, J, (Ed.) 1990. The London Corpus of spoken English: Description and Research. Lund Studies in English 82, Lund: Lund University Press.

Starvik, J. Ed. 1992. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 – Stockholm, 4-8 August 1991. Berling, New York, Mouton De Gruyer.

Thomas, J. and Short, M. (Eds.) 1996. Using Corpora for Language Research: Studies in the Honour of Geoffery Leech. London and New York: Addison Welsely Longman.

Tognini–Boneli. E.2001. Corpus Linguistics at work. Amsterdam: John Benjammins.

Umarani, P. 2001. Morphological Analyser for Tamil. Unpublished Doctoral Dissertation.

Vaishnavi Ramaswamy. 2000. A Morphological Generator for Tamil. M.Phil. Thesis, CALTS. University of Hyderabad, Hyderabad.

Veronis, T. (Ed.) 2000. Parallel Text Processing. Alignment and use of Translation Corpors. Dordreeht: Kluwer Academic Publishers.

Wichmann, A. Fligelstone, S. Mcenery, T. and Knowles, G (Eds.) 1997. Teaching and Language Corpora, London: Longman

Young, S. and G.Bloothooft (Eds.) 1997. Corpus based methods in Language and Speech Processing, Vol. II, Dordrecht: Kluwer Academic Press.